1993

Differential Comparison Standards and Their Subsequent Effects on the Agreement Between Self- And Supervisor Performance Appraisal Ratings.

Brian Wayne Schrader
Louisiana State University and Agricultural & Mechanical College

Order Number 9419924

Differential comparison standards and their subsequent effects on the agreement between self- and supervisor performance appraisal ratings

Schrader, Brian Wayne, Ph.D.

The Louisiana State University and Agricultural and Mechanical Col., 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

DIFFERENTIAL COMPARISON STANDARDS AND THEIR SUBSEQUENT EFFECTS ON THE AGREEMENT BETWEEN SELF- AND SUPERVISOR PERFORMANCE APPRAISAL RATINGS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Psychology

by
Brian Wayne Schrader
B.A., Bethany College, 1988
M.A., Louisiana State University, 1990
December 1993

ACKNOWLEDGMENTS

I wish to extend my sincere appreciation to my committee chair, Dr. Dirk Steiner, whose assistance, patience, and friendship was an invaluable asset in the completion of this dissertation. I would also like to thank Dr. Irving Lane, Dr. Stephen Gilliland, Dr. Katie Cherry, Dr. Susan Shackelford, and Dr. James Werbel for serving as members of my dissertation committee and for their contributions in the development of this research manuscript. A special thanks goes out to the various organizations in the Baton Rouge community which provided their time and services as subjects.

My gratitude is also extended to several of my friends who assisted me with various parts of this paper: Paul Damiano, Bruce Davis, and Stephen Lamoureux for coding the ratings, Drew Brock and Scott Klafke with data analysis assistance, and Mark Nagy and Beverly Andes for their help with the graduate school paperwork.

I would also like to my parents, Gloria Schrader and Robert Schrader, as well as my brother, Brad Schrader, who have always been there for me when I needed them. Their love, guidance, and friendship has been a constant blessing over the years.

Finally, I wish to express my love and appreciation for my wife, Angela, who supported me throughout the completion of my dissertation. Her endless encouragement, praise, and patience provided me with the emotional inspiration I needed to finish my graduate studies.

TABLE OF CONTENTS

(table continues)

TABLE OF CONTENTS (con'd)

LIST OF TABLES

(table continues)

LIST OF TABLES (con'd)

LIST OF FIGURES

ABSTRACT

This study examined differential comparison standards (i.e., comparative bases for performance evaluation) and their effects on the level of agreement between supervisory and self-raters (i.e., subordinates) within the context of a performance appraisal system. The purpose of the research was to determine whether differential comparison standards represented an underlying mechanism in the traditionally poor correlational relationship between self-supervisor performance ratings. Supervisor and subordinate rater dyads (N = 106 dyads) evaluated job performance across three dimensions using five different comparison standards (ambiguous, internal, absolute, relative, and multiple) in addition to providing preference, availability, and relevancy ratings. Results supported the hypotheses indicating that more explicit and objective comparison standards produced higher levels of interrater agreement, preference, availability, and relevancy. The implications of these findings are discussed, particularly in terms of comparison standards being adopted in current research and future performance appraisal systems.

INTRODUCTION

Performance appraisal systems have long been an important area of research in both academia and business. Historically, appraisal systems have centered around three key pieces of information in performance evaluation: objective production data, personnel data, and judgmental data. Of the three categories, judgmental data have the advantage of being readily accessible across a myriad of job types, can be obtained in a time- and cost-efficient manner, and have an extensive literature base of supportive research (Landy, 1989).

Judgmental data rely largely on subjective assessments of an individual's performance. Two popular approaches to obtaining judgmental data have been (1) supervisor ratings, where an employee's superior rates the employee across several performance dimensions, and (2) self-ratings, where the individual employee conducts an assessment of his/her own performance. Unfortunately, comparisons between the two rating approaches have resulted in conflicting and inconclusive findings across a variety of studies as to the true reliability and validity of the ratings (Fisher, 1989). Supportive and unsupportive research on the value of these rating approaches has continually attempted to isolate an underlying factor which produces these disparate findings.

However, there still exists considerable disagreement as to the source of the poor correlational findings between supervisor and self-ratings when conducted for performance appraisal purposes (Fisher, 1989). In an attempt to better understand the factor(s) underlying the disparity between rating sources, researchers have recently focused on different points of reference between the raters (e.g., Hauenstein & Foti, 1989). This line of reasoning asserts that self-raters approach the appraisal process from a different viewpoint and are influenced by different motivations than are their supervisors.

This subsequent discrepancy in the raters' viewpoints is often assumed to be the central mechanism which results in poor reliability and validity findings for self and supervisor rating comparisons. To overcome this discrepancy, frame-of-reference (FOR) training (e.g., Sulsky & Day, 1992), whereby both raters are taught similar performance dimensions and categories, may help to reduce the disagreement between the sources by providing raters with similar frames of reference.

Unfortunately, current research has neglected another potential source of disagreement in self- and supervisor ratings beyond what I will refer to as "differential reference points." An equally serious cause of rater disagreement which I will call "differential comparison

standards", reflects a difference in the reference groups (e.g., fellow co-workers) and/or standards (e.g., a specific or absolute goal) used by raters when seeking a comparative benchmark. Thus, differential reference points reflects a discrepancy in rater viewpoints as to the importance, weighting, and relevancy of various performance behaviors, whereas differential comparison standards represent the selection of distinctively different groups of people or standards to use as benchmarks for comparative purposes. Often, this difference in comparative referent groups is simply due to ambiguous wording encountered in performance appraisal instructions which fails to explicitly state the comparison standard to use (Landy, 1989).

This paper will argue that irrespective of the differential reference points problem, raters must know which comparison group to use when making appraisal evaluations in order to increase interrater agreement. Four potential differential comparison standards (internal, relative, absolute, multiple) exist for raters to choose from when conducting performance appraisals. Furthermore, it is believed that the conflicting self- and supervisor rating research findings are due to the use of differential comparison standards, which remain to be examined in the performance appraisal literature. The

present study seeks to provide support for the existence of these differential comparison standards, to explore how these standards are employed by the different raters, and to examine what their effects are upon correlations between self- and supervisory ratings.

LITERATURE REVIEW

Traditional Approach to Performance Appraisal

Performance appraisals have traditionally consisted of rating an employee's work performance by either objective, nonjudgmental measures (e.g., production output, completion time, number of errors) or subjective, judgmental evaluations. Whereas objective measures have their own unique strengths and weaknesses in the appraisal process, subjective appraisals are by far the more commonly employed technique (Murphy & Cleveland, 1991). A typical subjective performance appraisal generally consists of a single supervisor evaluating (i.e., rating) a subordinate on multiple performance criteria for a given job (Bernardin & Beatty, 1984). The use of supervisor-based evaluations has been well-documented in the literature, and they are a valid predictor of performance and ability as well as an established criterion in relation to other rating sources such as peers and objective data (Hunter & Hunter, 1984; Reilly & Chao, 1982).

Researchers have become increasingly discouraged with some significant problems inherent in supervisory ratings, including: susceptibility to rater biases (Cascio, 1987; Landy & Farr, 1980; Mount, 1984; Thornton, 1980), limited observational opportunities of subordinate's performance

(Heneman, Wexley, & Moore, 1987; Riggio & Cole, 1992), cognitive constraints (Campbell & Lee, 1988; DeNisi, Cafferty, & Meglino, 1984; DeNisi & Williams, 1988; Fisher, 1989), and substantially greater time and cost requirements compared to alternative measures such as self-ratings (Bassett & Meyer, 1968; Klimoski & London, 1974). Similarly, the context and purposes for which performance appraisals are used are also under investigation, suggesting that traditional approaches alone may not be optimal for effectively evaluating performance (Cleveland, Murphy, & Williams, 1989; Murphy & Cleveland, 1991). Lastly, a recent meta-analysis by Heneman (1986) found that supervisory ratings only correlated .27 with performance criteria (i.e., results-oriented measures) even after being corrected for sampling error and attenuation, indicating that traditional supervisory appraisals were far from perfect and may not have as much predictive validity as once thought.

Alternative approaches to performance appraisal. The shortcomings in supervisory ratings have led researchers to reexamine the entire performance appraisal process with a special emphasis on other rater types. One specific area of interest is the focus on how other inputs beyond the supervisor's may help to improve rater accuracy and provide additional evaluative information (Jones, 1991).

Two known alternatives are self-ratings and peer ratings. Self-ratings allow an employee to rate him- or herself on the same (or different) performance dimensions as the supervisor does (see Ashford, 1989 for a complete review of self-assessment processes). Peer ratings involve fellow co-workers from within the appropriate workgroup assessing the ratee across these same performance dimensions.

Typically, self- and peer ratings have been used in performance appraisal for three purposes: (1) as additional data points for a supervisor to consider (Campbell & Lee, 1988), (2) as an integral component of the appraisal process (Campbell & Lee, 1988), and/or (3) for developing employees by exposing their strengths and weaknesses (Steel & Ovalle, 1984). Additionally, self- and peer performance appraisals have been modified for detecting individual and organizational training deficits (Ford & Noe, 1987; McEnery & McEnery, 1987). However, peer appraisals are rarely used in employment contexts except in military settings and, as such, are limited in their applications to the performance appraisal process unless an appropriate pool of co-workers exists to provide observations (McEvoy & Buller, 1987). While peer appraisals can be an effective rater source, self-

appraisals retain the benefits of being less time consuming and more functional in dyadic relationships.

Self-Ratings Research

The performance appraisal literature has indicated both numerous advantages as well as disadvantages for including self-ratings in the appraisal process (Fisher, 1989). In addition, there has also been considerable disagreement as to the validity and reliability of self-ratings especially in direct comparison with other rating sources. As a result of the conflicting views towards self-ratings and their role in performance appraisal systems, researchers have been left with an issue which is divided and unresolved in terms of establishing a consensus. The following sections will explore the various literature and research which has left the current thought on self-ratings literature in a state of inconclusiveness.

Advantages of self-ratings. A large body of literature exists to support the usage and purported advantages of self-evaluations in the performance appraisal process, including: (1) increased user acceptance of the appraisal system due to subordinate participation (Latham & Wexley, 1981; Riggio & Cole, 1992; Shrauger & Osberg, 1981), (2) reduced defensiveness in the ratings (Farh, Werbel, & Bedeian, 1988; Latham & Wexley,

1981), (3) enhanced legal defensibility due to the use of multiple raters (Bernardin & Beatty, 1984), (4) increased observation of performance on relevant criteria (Borman, 1974; Henderson, 1984), (5) cost effectiveness in terms of time and money (Shrauger & Osberg, 1981), (6) enhanced relationships between supervisor and subordinates (Carroll & Schneier, 1982; Fletcher, 1986), (7) significant improvements in subordinate's performance following self-assessment (Bassett & Meyer, 1968), (8) increased communication between subordinates and supervisors resulting in less ambiguity in the appraisal process and improved resolution of rating disagreements (Fletcher, 1986), (9) less halo error than in supervisory evaluations (Thornton, 1980), and (10) a more comprehensive data base consisting of multiple ratings which can be used to make performance decisions (Carroll & Schneier, 1982). Additional advantages of self-ratings are abundant in areas outside the boundaries of performance appraisal such as training and job satisfaction (cf. Campbell & Lee, 1988; Cleveland et al., 1989; McEnery & McEnery, 1987; Thornton, 1980).

Many researchers also support the contention that self-ratings are the most appropriate, accurate, and valid assessment of performance because individuals are in the best position to evaluate their own work, especially in

situations where their need to inflate ratings is low (e.g., Fox & Dinur, 1988; Shrauger & Osberg, 1981). That is, since individual employees are privy to significantly greater amounts of performance information and feedback from multiple sources (i.e., self, peer, supervisor, task, company standards, etc.), they are more qualified to make inferences about their own abilities and performance than any other person.

Disadvantages of self-ratings. Despite the multitude of advantages, there has been a voluminous amount of opposing and/or conflicting research arguing that self-ratings are not effective in the performance appraisal process and are subject to a variety of psychometric problems. The most common limitation of self-appraisals has generally been considered their low agreement with other measures, which in turn, often leads to a general lack of convergent and discriminant validity (Fisher, 1989). A secondary consideration is their potential susceptibility to leniency on the part of the rater. Leniency and self-ratings have often been linked together under the basic premise that employees were psychologically pre-disposed to rate themselves high in regard to their work performance due to compensation considerations (Ashford, 1989).

Convergent and discriminant validity research. With regard to validity, self-appraisal research has varied considerably as to the amount of convergent (i.e., agreement among multiple sources) and discriminant validity (i.e., independence across multiple dimensions) evidenced in several studies.

Several studies have shown self-raters to exhibit at least moderate levels of agreement with other raters. A meta-analysis by Harris and Schaubroeck (1988) found self-appraisals to correlate .36 with peer appraisals and .35 with supervisor evaluations. A study by Fox and Dinur (1988) on predicting success over a 2-year period in military training found evidence of convergent validity between self-ratings and supervisory ratings and additional support of low, but significant correlations between self-ratings and both supervisor and peer rater sources. Likewise, Somers and Birnbaum (1991) provided support for convergent validity for self-appraisals with supervisory ratings using a multi-trait, multi-method approach with 8 of 10 performance dimensions significantly correlated. On the other hand, London and Wohlers (1991) found that self-ratings of supervisors produced greater discriminant validity than subordinate ratings of supervisors when using a multi-trait, multi-method approach to examine leadership and relationship issues in

an upward feedback study. Two additional studies have shown that knowledge of comparative information (i.e., knowledge of peer performance levels) can significantly increase correlations ($\underline{r}$ = .51 for overall evaluation) between self- and supervisory ratings of performance (Farh & Dobbins, 1989; Farh & Werbel, 1986).

However, there have been several studies which refute the supportive evidence presented above. In a comprehensive review of the literature, Landy and Farr (1980) concluded that a low to moderate relationship (at best) exists between multiple sources of ratings. Thornton (1980) indicated that, in general, ratings from different appraisal raters resulted in low intercorrelations and lacked discriminant validity. A meta-analysis by Mabe and West (1982) statistically confirmed the assumptions of the previous literature reviews in finding a mean correlation of .29 between self- and supervisory ratings. However, many have pointed to the considerable variation in the correlations (e.g., one study produced a -.26 correlation). Even the recent meta-analysis by Harris and Schaubroeck (1988) which produced a correlation of .35 between self- and supervisor ratings is overshadowed by a mean correlation of .22 when appropriately corrected for sampling error (Fisher, 1989). A previously mentioned study by Fox and Dinur (1988),

which was considered supportive of self-ratings because of their lower halo, was able to provide only low convergent validity and no evidence of discriminant or predictive validity for self-ratings. Similarly, whereas Steel and Ovalle (1984) could produce some evidence of convergent validity, there was no support for discriminant validity.

Predictive validity research. Predictive validity (i.e., the relationship between predictors and criteria) has also met with divided opinion in the literature as to the predictive abilities of self-ratings.

Mabe and West (1982) found a mean correlation of .29 between self-ratings and various performance criteria with 88% of the correlations greater than zero. Furthermore, the correlations were significantly higher ($\underline{r}$ = .64) when studies met more restrictive criteria (i.e., accounting for poor measurement conditions) and included moderators. An empirical review of self-ratings by Shrauger and Osberg (1981) indicated that self-appraisals were at least as accurate, if not better, than other performance predictors in the majority of studies. Self-ratings under more stringent methodological conditions (i.e., reliable criterion measures and increased variability in performance) had enhanced predictive power which increased over time and correlated with objective performance data

(r's ranged from .33 to .56), well beyond Mabe and West's (1982) level of .29 (Lane & Herriot, 1990).

On the other hand, reviews by Reilly and Chao (1982) and Hunter and Hunter (1984) of alternative predictors for performance and ability both discounted self-ratings as a valid predictor based on their low correlation with other predictors and criterion measures.

A recent empirical study by Hoffman, Nathan, and Holden (1991) compared self- and superior ratings to both objective and subjective performance criteria. Their results indicated that self-ratings had "near zero" validity with performance measures and produced low correlations with supervisory ratings.

Leniency and halo research. With respect to rater biases, Fox and Dinur (1988) indicated that self-ratings exhibited significantly less halo than other ratings. Studies by Farh and Werbel (1986) and Somers and Birnbaum (1991) both found self-ratings to be free from serious leniency error and the concomitant problem of range restriction under the more rigorous conditions identified by Mabe and West (1982)[1]. Somers and Birnbaum (1991)

---

[1]Mabe and West (1982) identified nine different measurement conditions which have come to be regarded as criteria for conducting self-rating research. The more criterion restrictions a study imposed, the more rigorous its methodology.

found correlations between self- and supervisor ratings ranging from .27 to .41. However, when the correlations were corrected for statistical artifacts and halo error, they reached an $\underline{r}$ = .64.

A study by Farh et al. (1988) on self-appraised performance evaluations produced several notable results. Incorporating a self-appraisal format into an existing traditional performance appraisal system for research purposes, Farh and associates used college faculty to explore the congruence between self- and supervisory ratings across a variety of performance dimensions (e.g., publications, departmental service, instructional method). They found no significant difference between the two rater types on leniency. In addition, correlations between performance criteria and various self-rating dimensions ranged from .37 to .63 which closely mirrored the correlations of supervisory ratings with performance criteria. Hence, self-ratings provided significant and strong support for convergent validity with supervisor evaluations. As an added bonus, user acceptance of the self-appraisal format was very high.

An empirical study, opposing the supposed lack of bias in self-ratings, was offered by Hoffman et al. (1991). The authors found that self-ratings were extremely prone to severe leniency. This recent finding is supportive of the

longstanding belief that self-ratings are vulnerable to the egos of the raters which use them (Landy & Farr, 1980; Reilly & Chao, 1982; Thornton, 1980). The resulting inflated ratings lead to mean rating differences between employee-supervisor dyads as well as poor agreement between multiple rating sources.

Conclusions. In summary, there exists a substantial amount of empirical evidence both for and against the use of self-ratings in performance appraisals. The reader should be left with the impression that no definitive conclusions can as yet be reached regarding the actual reliability and validity of self- and supervisory ratings. Interrater agreement between self-raters and supervisors has been shown to range anywhere from negative correlations to highly significant positive correlations, although the overall evidence suggests a weak positive correlation between the two rating sources. Similarly, reports on convergent and discriminant validity have fluctuated between both ends of the continuum. Thus, a synopsis of the performance appraisal literature suggests that previous research findings are inconclusive and have failed to adequately explore both existing and proposed methods for resolving the discrepant findings. More recently, however, researchers have begun to examine possible explanations for the cause of the conflicting

studies (e.g, Fisher, 1989) in an effort to uncover the source of these equivocal findings.

Underlying Problems in the Self-Rating Literature

Taking both supportive and opposing research into account, the extensive literature on self-ratings in performance appraisal suggests that some underlying mechanism may exist to account for the conflicting and inconclusive results on multiple rating sources. There have been several proposals made within the self-rating literature regarding the true source of the discrepancy. The vast majority of these hypotheses fall within one or more of four categories: (1) rater error and rater biases, (2) actor/observer differences, (3) political influences, and (4) cognitive and informational constraints. Each of these categories will be considered in the following sections.

Rater error and rater biases. Leniency has been, by far, the most widely cited psychometric problem with self-appraisals. Leniency error occurs when individuals systematically rate themselves higher, on average, across multiple dimensions when compared to other rating sources. Numerous studies have empirically demonstrated the contention that leniency error is a serious threat to self-rating validity by finding a significant difference between group means across dimensions for multiple rating

sources (Farh & Werbel, 1986; Fox & Dinur, 1988; Hoffman et al., 1991; Klimoski & London, 1974; Mabe & West, 1982; McEnery & McEnery, 1987; Meyer, 1980; Parker, Taylor, Barrett, & Martins, 1959; Shrauger & Osberg, 1981; Steel & Ovalle, 1984; Thornton, 1968, 1980). Hence, individuals evaluating themselves on performance criteria tend to inflate their ratings relative to peer or supervisory ratings of that same individual. Although leniency error alone does not conclusively convict self-ratings of invalidity, its close relationships with range restriction, negatively skewed distributions, and variability reduction exhibited in self-ratings are problematic (Murphy & Cleveland, 1991). Both the restriction of range and limited variability weaknesses in connection with leniency have been well-documented in the self-appraisal literature (Fisher, 1989; McEnery & McEnery, 1987; Thornton, 1980).

Another rater bias often connected with self-ratings is halo. As Balzer and Sulsky (1992) propose, halo can occur in one of two forms, either (a) General Impression Halo, when a rater generates an overall impression toward a ratee and this impression consistently biases the rater's evaluation of the ratee or (b) Dimensional Similarity Halo, when a rater perceives high intercorrelations among performance dimensions and thus

rates an individual similarly across like dimensions. Because these operational definitions have only been proposed recently, it is often unclear as to which type of halo was examined in previous studies. Nonetheless, self-appraisal studies have been notorious for claiming that minimal halo exists in self-ratings (e.g., Thornton, 1980). Nathan and Tippins (1990) produced evidence that the presence of halo actually results in higher validity findings for ratings. Hence, the lack of halo in self-ratings may be partially responsible for low correlations with supervisory ratings. However, Balzer and Sulsky (1992) advocate caution in interpreting halo findings in performance appraisal research since halo can have a positive, negative, or zero effect on rater accuracy and recommend that halo not be used as a consideration of rating validity.

Actor/observer differences. A second potential reason behind the conflicting self- and supervisor ratings' literature may be derived from differing attributional processes in the two raters. Jones and Nisbett (1971, 1972) termed these opposing attributional perspectives as "actor-observer differences." In essence, individuals performing in an ambiguous situation (i.e., actors) are likely to attribute their own behavior to external causes (i.e., luck and situational constraints), whereas

observers are likely to make internal attributions (i.e., effort and ability) about others' performance. In less ambiguous circumstances, such as a structured work setting, individuals display a tendency to alter their attributions to match the success or failure of the performance (Fisher, 1989; Weiner, 1986). Actors make internal attributions for successful performance on the job and external attributions for failure (Gioia & Sims, 1986). Observers, on the other hand, make external attributions for successes of the actor and internal attributions for failures. Obviously, the terms self and supervisor could be substituted for actor and observer in the rating context. A recent study by Arnold and Davey (1992) found that graduates entering a new job were more inclined to make internal attributions for success than their supervisors, who were more likely to make external attributions, as evidenced by a comparison of self and managerial ratings. Harris and Schaubroeck (1988) explained their levels of agreement between self, peer, and supervisory ratings in relation to actor-observer differences, arguing that the reason peer and supervisor ratings were the most highly correlated (r = .62) was due to both of the rating sources being "observers" who used external attributional processes for successful performance. Self-rating correlations with peers (r =

.36) and supervisors (r = .35) both contained one actor and one observer who generated their performance ratings from different perspectives, leading to significantly lower interrater agreement. The implications of this analysis are that other studies which found poor self-rating correlations with supervisors may have been affected by opposing attributional processes.

Political influences. The effects of political interplay between raters may also have considerable impact on subordinate-supervisor ratings. The most prominent line of research in this area focuses on self-enhancement tactics employed by the subordinate. Many researchers have suggested that the underlying reasoning behind increased leniency on the part of subordinates is their desire to appear competent and successful to their supervisor (Ashford, 1989; Fisher, 1989; Murphy & Cleveland, 1991). Hence, many employees will tend to inflate their self-appraisal ratings to look good in the eyes of their superiors. Meanwhile, supervisors may increase or decrease subordinate ratings of performance to meet their own special needs (Longenecker, Sims, & Gioia, 1986). For example, in an effort to punish a rebellious or troublemaking employee, a supervisor may intentionally give poorer marks than are reflective of the subordinate's true performance. Conversely, a supervisor may inflate

ratings to reward employees or increase their chances of promotion (possibly even to incompetent performers). The executives who participated in the Longenecker et al. (1986) study also indicated that accuracy in performance appraisals was not nearly as important as affecting future performance in individuals and the workgroup. The resulting effect may be a failure for either self- or supervisory ratings to be truly representative of the subordinate's actual abilities and performance across dimensions (Campbell & Lee, 1988). Such a predicament would undoubtedly lead to reduced correlations between rating sources and eliminate the likelihood of finding convergent validity (Fisher, 1989). Thus, the inescapable realities of a socially constructed, political organization are likely to have significant effects on the actual ratings between self- and supervisory raters within the performance appraisal system.

Cognitive/informational constraints. The fourth potential mechanism underlying discrepancies between different raters may be due to cognitive and/or informational constraints on the rater. DeNisi, Cafferty, and Meglino (1984) illustrated the basic cognitive processes which occur during performance appraisal. The process begins with observation of the specific job performances, followed by formation and storage of the

cognitive representation, retrieval of the representation for evaluation purposes, reconsideration and integration with other knowledge, and finally evaluation. Due to the complexity of the entire process, raters are forced to rely on cognitive categorizations or schemas (Ilgen & Feldman, 1983). Schemas are used to classify information quickly about various stimuli as well as to develop expectations, attributions, and spatial-temporal relationships in reference to the stimuli. However, people are often limited and/or inaccurate in their ability to recall schemas completely. Unfortunately, there is evidence to suggest that subordinates and superiors may possess qualitatively different schemas of performance (Bernardin & Beatty, 1984; Fisher, 1989). Furthermore, their ability to encode and retrieve information is subject to a variety of individual differences. The overall result of these processes suggests that self-ratings are likely to be markedly discrepant from ratings obtained from other rating sources (Bernardin & Villanova, 1986; McEnery & McEnery, 1987). Whereas a complete review of the cognitive literature is beyond the scope of this paper (see Fisher, 1989 for a thorough review of cognitive schemas in self-appraisal research), it is sufficient to note that the impact of

cognitive factors on multiple raters is clearly a probable reason for interrater disagreement.

Informational constraints may also reflect real differences in rater agreement. Obviously, self-raters have much more access to knowledge of their performance, especially on a day-to-day basis. Supervisors, in general, have fewer observational opportunities and less spare attention to devote to individual employees (Fisher, 1989; Ilgen & Feldman, 1983). Furthermore, supervisors may have inaccurate or incomplete knowledge about the true nature of the subordinate's job. The problem of informational differences may be further compounded when the job/task or the work environment is relatively ambiguous (Ashford, 1989; Campbell & Lee, 1988). Such circumstances are likely to prevent adequate feedback opportunities for either rater. In sum, both cognitive and informational constraints pose considerable difficulty in establishing convergent validity between self- and supervisory ratings.

Differential reference points. Taken together, it should be apparent that the four potential mechanisms (rater biases, actor-observer differences, political reasons, and cognitive/informational constraints) which underlie the discrepancies between self- and supervisory performance ratings are interrelated. That is, each of

the four underlying problems either directly state or indirectly imply that the raters are approaching the rating process from significantly different points of view. Whereas the actual theoretical underpinnings may differ, it is clear that the four mechanisms are represented by this similar theme. The considerable overlap among these four problems allows for combining the mechanisms into a single complex problem in the self-rating literature: differential reference points. Namely, raters of all sources (i.e., self, peer, subordinate, and superior) are essentially entering into the performance appraisal process from different perspectives or frames of reference (Borman, 1974; Fisher, 1989; Klimoski & London, 1974). That is, raters are approaching the appraisal process with disparate reference points; self-raters are more likely to be lenient, make internal attributions for success, inflate ratings for self-enhancement, and have more access to self information, whereas supervisors rating employees are less lenient, make external attributions for success, alter ratings as dictated by their needs, and have greater cognitive demands with more informational limitations. Hence, self-raters and their supervisors approach the performance appraisal process from markedly different vantage points. Subsequently, it is not surprising that interrater agreement between

multiple rating sources has suffered from weak correlations in the literature. Often, researchers have responded by investigating the effects of rater training on rater accuracy. Early rater training systems such as those advocated by Pulakos (1984) generally focused on either increasing accuracy or decreasing errors. Unfortunately, both methods emphasized the importance of halo and leniency rather than providing raters with more similar frames of reference. More recently, however, frame-of-reference (FOR) training has been shown to be an effective framework for illustrating how multiple rating sources could increase rating accuracy through the use of shared reference points (Athey & McIntyre, 1987).

Frame-of-reference training. Frame-of-reference (FOR) training is a relatively recent advancement for improving rater accuracy (Hauenstein & Foti, 1989; Pulakos, 1984). The basic tenet of FOR training is to standardize raters' conceptions and perceptions of performance (and dimensions) so that raters will have a similar reference point (i.e., prototype) (Athey & McIntyre, 1987). Consequently, FOR training appears to be capable of compensating for the difficulties generated by differential reference points. McDonald (1991) found that when raters where given similar frames of reference and information regarding performance dimensions, rater

attentional processes improved, but, more importantly, rater accuracy increased to the point of being comparable with expert raters. Sulsky and Day (1992) found that FOR-trained raters have enhanced classification accuracy (recalling whether someone is a good or bad employee) but poor behavioral accuracy (recalling whether individuals performed specific behaviors or not). This lack of behavioral accuracy, of course, could be problematic for performance appraisal ratings which tend to focus on evaluating employees across a broad range of performance dimensions which consist of numerous behaviors. Fisher (1989) has called for subordinates to be trained in rater accuracy and knowledge of performance dimensions in an effort to improve rater agreement via FOR training. Despite the limiting problem of behavioral accuracy, FOR training seems to present a viable approach to reducing differential reference points; however, more research is needed.

One area which remains to be investigated either independently or in conjunction with the FOR training rubric is how various raters select the benchmarks or standards on which to base their ratings. That is, irrespective of the discrepant points of reference problem amongst rating sources, there is an additional need to

explore differences in the raters' selection of comparison groups on which to base their performance standards.

Differential Comparison Standards

Although FOR training shows much promise in reducing the problems imposed by disparate frames of reference, differential reference points (created by the four sources: rater biases, actor-observer differences, political reasons, and cognitive/informational constraints) are fundamentally distinct from a second source of rater disagreement which I refer to as differential comparison standards. Whereas the former category has been extensively researched and documented, significantly less research has been conducted on differential comparison standards and their effects on performance appraisal ratings.

A **comparison standard** can be defined as a particular referent choice which serves as the presiding benchmark on which performance comparisons are based. Differential comparison standards occur when raters select different comparative referent individuals, groups, and/or specific standards on which to base their ratings. For instance, self-raters may prefer to base their performance ratings on their own personal, internal standards. Alternatively, raters might wish to base their ratings on known company standards or perhaps on comparisons to other co-workers.

Each of these referent choices represents a unique comparative standard. Thus, not only may multiple raters approach the rating process from different frames of reference, they may also be using different standards of comparison when evaluating themselves or others on the various performance dimensions. Some research exists to support this proposal.

Steel and Ovalle (1984) found that when raters were allowed access to performance appraisal feedback so as to create a shared comparison reference, correlations between self and supervisory ratings increased. Similarly, meta-analyses by Mabe and West (1982) and Heneman (1986), as well as a study by Farh et al. (1988), found that the magnitude of correlations between multiple sources of ratings significantly increased when comparative instructions or information was given providing common standards. A more recent study by Farh and Dobbins (1989) examined the extent to which self-ratings correlated with objective performance measures and supervisory ratings when subjects were provided with differing amounts of comparison information on their co-workers. Results indicated that subjects who were exposed to comparative data produced ratings which were more highly correlated with both objective measures and supervisory ratings than control subjects who had no comparative exposure. Their

findings suggest that when supervisors and employees have a shared comparison standard to evaluate their performance, interrater agreement increases between supervisor and self-appraisals. Research by Summers and DeNisi (1990) allowed raters the selection of nine different referent choices such as others within the company and others with the same job title. Their results indicated considerable variability in referent choice, and the authors concluded that the availability of multiple referent groups was an important issue in understanding referent selection. McEnery and McEnery (1987) found that supervisors were employing categorically different comparison standards than their subordinates. Managers appeared to be using personal, internal standards since their ratings of subordinates were significantly correlated with the manager's <u>own</u> training needs. Fisher (1989) suggested that supervisors may in fact, "... use their own past or imagined performance in the subordinate's job as a standard against which to evaluate subordinates" (p. 46). Stepina and Perrewe (1991) investigated comparative referent choice under conditions of inequity and found that whereas many individuals used only a single comparison standard for compensation, different standards were used for other job facets. In addition, these comparative standards were unstable and

often changed over time. The implication was that raters are likely to draw on different comparison groups for performance dimensions and that these comparison standards may change over time. However, while the literature has generally supported the notion of comparison standards, most of the research has only tangentially explored the possibility of comparison standards as a major source of low agreement between self- and supervisory ratings.

Whereas there have apparently been no studies that have empirically investigated differential comparison standards used by raters in the performance appraisal context, there are several psychological theories which lend credence to the existence and importance of differential comparison standards. Theoretical considerations include: (1) equity theory, (2) social comparison theory, and (3) relative deprivation principle/theory.

Equity theory. Adams' (1963, 1965) equity theory proposes that individuals generate a ratio of inputs (e.g., performance on the job, education, training, work experience, etc.) to outcomes (e.g., pay, benefits, job security, etc.). Adams suggests that people differentially weight these inputs and outcomes with respect to their importance and relevancy. They then compare their input-outcome ratios to those of other

individuals in their surroundings. The comparison "others" could be co-workers, supervisors, subordinates, or some third party. Equity is said to exist when an individual perceives the ratio to be equal to the ratio of the comparison other. Inequity exists when the ratios are unequal.

Although the bulk of equity theory research has focused on reactions to compensation equity/inequity in an employee-employer exchange process, there has been some attention to the selection of comparison standards (Mowday, 1987). Goodman (1974) listed three referent classes: (1) others, (2) self-standards, and (3) system referents as possible comparison standards. "Other" referents could be further classified as "other-inside" (i.e., persons within the same work organization) or "other-outside" (persons outside the organization). System referents were explicit or implicit contractual requirements between employee and employer (i.e., external standards). Stepina and Perrewe (1991) found that not only did employees use multiple reference standards within an equity framework, but that these comparative referents were subject to change over time in many circumstances. Summers and DeNisi (1990) used equity theory to further

explore Goodman's three classes of referent[2]. Although their study focused on pay equity, the authors found that over 34% of subjects relied on self-standards, 20% used other-inside, almost 6% used other-outside, and over 37% used some form of generalized comparison standard (i.e., external sources or combinations of the other three). A similar study by Dornstein (1989) investigated referent comparisons with regard to pay in an equity framework. He found that individuals do in fact consider coworker comparison groups when determining compensation equity.

Social comparison theory. Festinger's (1954) social comparison theory also provides theoretical support for differential comparison standards. According to social comparison theory, people desire to obtain stable and accurate assessments of their own personal abilities. Oftentimes, this is accomplished by simple comparison with some existing objective measure (e.g., running a 4-minute mile, reaching a sales quota, getting a 94 on a history exam). However, in the presence of more "ambiguous" objective standards where individuals cannot rely on self-assessment or comparison to a known objective measure, they will compare themselves to other individuals. This

---

[2]This study excluded the system referent because the response format did not allow for it.

comparison may take many forms including: self-equality (comparing oneself to someone who is perceived to have equal ability), self-enhancement (comparing oneself to someone who is perceived as having less ability), and self-depreciation (comparing oneself to someone who is perceived to have more ability) (Levine & Moreland, 1986, 1987). Given a choice, self-raters seem to prefer self-enhancement when evaluating their abilities (Fisher, 1989). This may account for the tendency towards greater leniency in self-assessments. However, superiors are not likely to succumb to this self-enhancement motivation since they are not rating themselves but rather an employee. Fisher (1989) states, "Clearly, if superior and subordinate are using different comparison others, they are likely to reach different conclusions and disagree about the level of subordinate performance" (p. 23). Using the assumptions inherent in social comparison theory, if multiple rating sources were "forced" to use the same comparison standard, interrater agreement should increase. This hypothesis has been indirectly supported in the work of Mabe and West (1982) who found that self-superior correlations and agreement with objective performance measures were higher when ratings were made on a relative scale (i.e., compared to other individuals) as opposed to an absolute scale (i.e., compared to an

established goal level). Farh and Dobbins (1989) also worked within a social comparison framework and saw interrater agreement increase when raters were given the opportunity to observe all co-workers on specific performance dimensions than when comparative information was denied to the raters. Kruglanski and Mayseless (1990) presented some limitations on existing social comparison theory pointing to its narrow scope of focus. The authors indicated that the social comparison phenomenon may rely more on complex motivational processes and information accessibility rather than the rater consciously selecting a referent group. Thus, the selection of comparison standards is subject to wide variability across situations. More recent research in social comparison theory has begun to explore some of these underlying motivational and informational processes in selecting comparison standards (Suls & Wills, 1991).

Relative deprivation principle/theory. Relative deprivation theory is closely related to social comparison theory. The relative deprivation principle proposes that an individual's sense of happiness and satisfaction is tied to one's current perception of how one stands in relation to others in the environment (Myers, 1992, p. 401). However, while social comparison theory focuses on perceptions of ability, relative deprivation theory

emphasizes more material comparisons. An employee making $50,000 a year will feel happy and satisfied if fellow co-workers earn well below that income level and that is who the employee compares him/herself to. However, that same employee would be very unhappy and dissatisfied if his peers all earned in excess of $60,000. However, there appears to be a natural tendency for people to feel worse off than comparative others because we tend to compare ourselves, in terms of our possessions, to people better off than we are (Myers, 1992). Thus, perceptions of our relative standing with our peers influence our self-ratings and often in a self-deprecating manner. An empirical study by Sweeney, McFarlin, and Inderrieden (1990) found that satisfaction with current pay levels decreased when the similarity of co-workers increased. That is, employees were content when making significantly more than their peers, but as this compensation gap narrowed, contentment with pay plummeted. Although the absolute level of pay remained the same, one's sense of relative deprivation altered one's perceived happiness. This line of research would suggest that not only are individuals likely to possess different comparison standards, these referents of choice are influenced by the current situation. Logically then, relative deprivation is likely to be a factor in the performance appraisal

process where subordinates and supervisors are continually making judgments about the performance of others.

Levine and Moreland (1987) and Oldham et al. (1986) both argued that the process employed by individuals to decide on which comparison standard to use within a relative deprivation framework is primarily driven by the availability and relevance of the standard, where availability represented the accessibility of referent information and relevance represented the situational importance of the information. Furthermore, research suggests that employees are more likely to select an internal (i.e., self) referent or relative (i.e., workgroup) standard since they are generally available and relevant, whereas supervisors (who are dissimilar to the rest of the workgroup) are more apt to employ non-relative standards since intergroup comparisons may not be considered relevant even if they are available (Kulik & Ambrose, 1992; Oldham et al., 1986). Therefore, an employee's perceptions of his/her current status within the organization, department, and/or workgroup as to the relevance and availability of performance feedback is likely to affect the choice of a comparison standard.

Classification of comparison standards. Recently, Kulik and Ambrose (1992) have proposed that all three comparison theories (i.e., equity, social comparison, and

relative deprivation) are compatible and work in conjunction to create differential comparison standards. However, the authors argue that all three fail to identify which comparison referent group is used and how an individual arrives at that decision. They present a general model that incorporates all three theories as well as the mediating concepts of referent relevance and information availability to explain referent choice selection. In their proposed framework, Kulik and Ambrose examined the effects of a variety of personal and situational determinants on referent selection. One particularly interesting finding was drawn from the work of Oldham, Kulik, Ambrose, Stepina, and Brand (1986) who noted that, given a choice, people relied on self-referents (i.e., using their own personal standards) over 56% of the time. Kulik and Ambrose (1992) suggested that individuals may use their own personal, internal values as their comparison standard of choice and were likely to use it as a "default" referent choice in situations involving limited or ambiguous information. Other comparison standards would only be considered when they were deemed relevant and possessed comparative information.

Despite the supportive theory and research, there have been no studies directly investigating differential comparison standards in the performance appraisal

literature. It has already been suggested that including comparative data when giving ratings may serve to improve ratings from multiple sources (Farh & Dobbins, 1989; Mabe & West, 1982; Steel & Ovalle, 1984). However, Farh and Dobbins (1989), Mabe and West (1982), and Steel and Ovalle (1984) all manipulated the extent to which comparative data were available to the subject. Unfortunately, while this manipulation is relatively easy in a laboratory environment, the controlled restriction or inclusion of comparative data is unrealistic in organizational settings. That is, with a few possible exceptions, all individuals are privy to comparative data within their immediate workgroup. Most jobs also allow for comparisons beyond the immediate workgroup (e.g., professional athletes can compare themselves to teammates and/or players on other teams; secretaries can compare themselves to others within the office and/or to secretaries in other departments). Additionally, many jobs allow for comparison to external objective standards (e.g., producing X amount of widgets in Y amount of time; typing 60 words a minute). Finally, individuals can use their own personal, internal standards to evaluate their performance (e.g., being timely and efficient with daily paperwork). Therefore, most individuals have access to three different comparison standards: (1) internal (i.e.,

comparison to self), (2) **relative** (i.e., comparison to others), and (3) **absolute** (i.e., comparison to some objective measure). These standards are analogous to Goodman's (1974) referent typology. However, Goodman's referent typology failed to include the possibility that selection of comparison standards was a complex process wherein raters may combine aspects of each referent group to arrive at final standard. Thus, to extend Goodman's typology, a fourth possible comparison standard could be represented as a combination of the first three with a rater essentially drawing evaluative information from all three standards. This fourth referent choice is considered a **multiple** standard.

THE PRESENT STUDY

Based on the above propositions and supportive theoretical literature, this study will argue that the major underlying mechanism behind the disagreement in self- and supervisor performance ratings is in fact due to superiors employing a different comparison standard than self-raters. For example, self-raters may prefer using an internal or multiple standard whereas supervisors may employ a relative or absolute standard. Obviously, numerous comparison standard combinations (e.g., self-absolute, superior-internal) exist for any given self-superior rating pair. It is further hypothesized that if the use of comparison standards is not discussed prior to evaluations and/or performance appraisal rating formats are not specific in their instructions as to which comparison standard(s) is(are) to be considered, subordinate and supervisory ratings are likely to have low interrater agreement consistent with previous studies. That is, if the comparison standard of choice (e.g., relative) is not explicitly articulated either in the appraisal instructions or a pre-rating briefing to both raters, then they are not likely to select the same comparison standard.

The present study examines the effects of differential comparison standards on self- and supervisory ratings of

performance by providing raters with shared comparison standards. It is suggested that discrepancies in ratings from multiple sources is a function of which comparative standard each individual rater is employing. Furthermore, if rating formats are not specific in indicating which comparative standard the rater is supposed to be assessing, the resulting weak correlations will be due to ambiguous rating instructions.

Based on the literature, there should be a significantly higher correlation between self- and supervisory ratings when both raters are instructed as to which comparative standard is to be considered when conducting the performance appraisal. For example, using a simple 9 point Likert scale with 1 being the poorest rating and 9 being the best, instructions for each of the different rating standards might appear as follows:

Ambiguous - "Rate employee X on typing ability."

Absolute - "Rate employee X on typing ability in reference to the minimum acceptable standard of 60 words per minute."

Relative - "Rate employee X on typing ability compared to all other typists in your workgroup."

Internal - "Rate employee X on typing ability in reference to his/her own past performance and utilization of his/her individual skills.

Multiple - Rate employee X on typing ability considering all available sources of performance with respect to minimum company standards, comparison to coworkers, and individual ability.

The conditions under which using similar standards should improve agreement between self- and supervisory raters are such that, (a) performance is free to vary, (b) information about the aspects of the individual's internal standards, such as past performance and/or abilities, is available to supervisors, (c) past performance information for the workgroup is available for comparative purposes, (d) absolute standards (i.e., the expected minimum or average objective performance measures via company policy) are explicit and performance relative to them is available, and (e) appropriate coworkers exist on which to base comparative information. Although difficult, strict adherence to these conditions in the field study will result in enhanced internal validity (Cook & Campbell, 1979).

Hypotheses

Based on the self and supervisory performance appraisal rating literature and the above assumptions, the following hypotheses will be considered:

Hypothesis 1. The performance ratings of self-raters as well as supervisory raters will significantly differ as

a function of which comparison standard (ambiguous, internal, absolute, relative, and multiple) is employed by the rater across all three performance dimensions.

Confirmation of this hypothesis will provide supportive evidence for the existence of differential comparison standards and a potential underlying cause of self-supervisor disagreement. That is, support of the hypothesis will indicate that raters are in fact providing significantly different ratings dependent on the comparison standard stated in the instructions.

Hypothesis 2a. Interrater agreement between self- and supervisor raters, when collapsed across the three performance dimensions, will be greater for the explicit comparison standards (absolute, relative, internal, and multiple) than for the ambiguous comparison standard (which does not provide the rater with specific comparison instructions).

Hypothesis 2b. Interrater agreement between self- and supervisor raters, when collapsed across the three performance dimensions, will be greater than the previous self-supervisory correlations in the performance appraisal literature for the explicit comparison standards[3].

---

[3]Harris & Schaubroeck (1988) r = .35 for self-supervisor ratings.

Hypothesis 2c. Interrater agreement between self- and supervisor for the ambiguous comparison standard, when collapsed across the three performance dimensions, will not significantly differ from previous self-supervisory correlations in the performance appraisal literature.

The remaining hypotheses will examine which comparative standards are preferred by raters and how raters select their standards. The first of these remaining hypotheses investigates which referent group raters will select, prior to providing ratings, when given the opportunity to freely respond without being prompted by the explicit comparison standard alternatives.

Hypothesis 3a: Self-raters will prefer a comparative referent standard which is operationally equivalent to the internal comparison standard (i.e., self-referent) when asked in an open-ended format.

Hypothesis 3b: Supervisory raters will prefer a comparative referent standard which is operationally equivalent to the multiple comparison standard (i.e., combination of several referents) when asked in an open-ended format.

Hypothesis 3a is based on the rationale that employees prefer to use an internal comparison standard as evidenced in the research by Oldham et al. (1986) and Kulik and Ambrose (1992). Alternately, Hypothesis 3b is linked to

Longenecker et al. (1987) which found that supervisors are more likely to prefer a combination of factors based on the competing demands inherent within the performance appraisal process.

The next hypotheses will examine rater preference when given the opportunity to choose their preferred source from internal, absolute, relative, and multiple comparison standard choices upon completion of the performance appraisal ratings.

Hypothesis 4a: Both self- and supervisory raters will prefer to utilize the **multiple** standard when asked to rate each of the four explicit comparative standards (internal, absolute, relative, and multiple).

Hypothesis 4b. Self-raters will prefer the **internal** standard (after the multiple standard) for future performance appraisals, followed by the absolute and relative standards respectively, when asked to rate each of the four explicit comparative standards.

Hypothesis 4c. Supervisory raters will prefer the **absolute** standard (after the multiple standard) for future performance appraisals, followed by the relative and internal standards respectively, when asked to rate each of the four explicit comparative standards.

The rationale for Hypotheses 4a, 4b, and 4c are similarly linked to the work done by Oldham et al. (1986),

Kulik and Ambrose (1992), and Longenecker et al. (1987). Hypothesis 4a argues that both self-raters and supervisors will prefer the multiple comparison standard, since it incorporates more feedback allowing for a more comprehensive evaluation. However, Hypothesis 4c posits that supervisors will prefer an absolute comparison standard as a secondary choice since their position is tied to maintaining specific performance goals in their subordinates.

The last hypothesis will examine how raters determine which comparative standard to use based on ratings of availability and relevancy as proposed by Kulik and Ambrose (1992). It is anticipated that more available referents and more relevant referents will tend to receive higher ratings by both self- and supervisory raters. Hence, it is surmised that a rater's comparison standard selection process is guided by high levels of relevant and available performance information. In addition, greater levels of availability and relevancy are also more likely to produce higher preference ratings.

Hypothesis 5. The preference ratings of self-raters as well as supervisory raters for the four explicit comparison standards (internal, absolute, relative, and multiple) will significantly differ as a function of the availability and relevancy of the comparison standard.

METHOD

Subjects

The research was conducted using supervisors and subordinates (i.e., self-raters) across nine different organizations in a large Southern city. The organizations consisted primarily of financial institutions and retail department stores but also included a post office, a telemarketing firm, and a cosmetics outlet. Supervisors and subordinates represented a variety of job types ranging from branch managers and department managers to bank tellers and sales associates. Table 1 presents a complete breakdown of the participating organizations used in the study while Table 2 lists the job types by rater source.

An initial total of 162 rating pairs (supervisor-subordinate dyads) were available for the study. The use of a rating pair presupposed the presence of at least three other subordinates with job types similar to the self-rater. In addition, the other subordinates had to be directly supervised and evaluated by the supervisor. This was necessary to facilitate the comparative referent group for the relative comparison standard. Only subjects (supervisors and subordinates) who had been employed at their present job for at least six months were used in the sample pool.

Table 1

Listing of Organizational Types

ORGANIZATIONAL TYPE	n	SUBJECT n
Financial institutions	4	128
Retail department stores	2	54
Post offices	1	16
Retail cosmetics companies	1	10
Telemarketing firms	1	4
	N = 9	N = 212

Table 2

Listing of Supervisor-Subordinate Job Types

JOB TYPE	n
SUPERVISORS	
Branch Managers	64
Department Managers	27
Customer Service Managers	8
Counter Managers	5
Team Leaders (Managers)	2
	N = 106

JOB TYPE	n
SUBORDINATES	
Bank Teller	64
Retail Sales Associate	27
Postal Carriers	8
Cosmetic Consultants	5
Operators	2
	N = 106

Of the 324 packets (162 supervisors and 162 subordinates) issued to the organizations, 243 were completed and returned for a response rate of 75 percent. However, an additional 31 packets were not valid because the packets either did not represent a complete supervisor-subordinate dyad (n = 16) or the subject indicated that the packet instructions had not been fully understood (n = 15). Thus, a final total of 212 packets (n = 106 for both supervisors and subordinates) were available for analysis.

Subject's ages ranged from 18.0 to 74.0 years old with an average supervisor's age of 37.4 while the mean subordinate's age was 30.4. Fifty of the participants were male, 162 were female. The average number of years supervisors had worked at their present job was 5.3 and the mean number of years with their current company was 11.9. For subordinate's, the average number of years at their present job was 4.2 and 5.5 years with their current company. Finally, the average number of subordinates under a supervisor's direction was 8.8.

Procedure

Subjects' packets included a series of performance appraisal rating sheets, rankings, and rating-related questions. The front page of the packet contained instructions for the subjects in addition to serving as an

informed consent sheet (see APPENDIX A). Each subject answered a question on the use of comparative standards prior to the self- and supervisory evaluations. Each subordinate (i.e., non-supervisor) then provided self-ratings on a series of performance appraisal evaluation sheets aimed at assessing performance across three dimensions over the past six months. Two performance dimensions were used which were reflective of the organization's actual performance dimensions. A third dimension assessed overall performance.

The rating sheets only differed in their instructions to the rater as to which comparison standard the rater should consider when issuing the ratings. The subordinate's supervisor provided supervisory ratings for the subordinate using identical performance appraisal rating sheets. Next, all subjects were asked to provide preference ratings for each of the four explicit comparison standards. Finally, raters were asked to indicate the degree to which each of the four explicit comparative standards was available and relevant to the individual within their own job context.

Measures

Pre-rating comparison standard question. Each subordinate was asked to answer the following question prior to conducting the performance appraisal ratings,

"Please think about how you would rate your own job performance. If asked to evaluate your own performance on the job (i.e., provide a self-rating) what would you use as the basis for your ratings? That is, how would you decide whether or not you were performing satisfactorily on the job?". Each supervisor was asked a similar question, "Please think about how you would (or do) rate your employee's job performance. If asked to rate an employee on his/her job performance (i.e., providing a supervisory rating), what would you use as the basis for your ratings? That is, how would you decide whether or not the employee was performing satisfactorily on the job?".

Subjects were allowed to answer the question in an open-ended format (see APPENDIX B). This format was used to allow free response in subject answers as opposed to traditional forced choice alternatives. Responses were then coded by three subject matter experts (graduate students in Industrial/Organizational Psychology) into one of the four comparison standard categories (internal, relative, absolute, or multiple) or two additional categories which subjects employed but did not conform to the prescribed comparison standard categories. These two additional categories represented an "Other" comparison standard and a "N/A" (non-applicable) category for

subjects who either left the question blank or responded inappropriately.

The "Other" comparison standard was operationalized as a basis for performance ratings which incorporated subjective assessments either in isolation or in some combination of subjective assessments with the explicit comparison standards. Typically, responses in this category included subjective assessments such as: appearance, motivation, positive attitude, and communication skills. Responses which included even one of these subjective factors were classified as "Other" even if some of the explicit comparison standards were also included in the response.

Interrater agreement for the response coding done by the three subject matter raters was calculated as total agreement expressed as a percentage. The percentage of agreement for each rating pair was as follows: Rater1 - Rater2 (89.2%), Rater1 - Rater3 (89.6%), and Rater2 - Rater3 (90.6%).

Self-evaluations. Each subordinate was asked to make self-ratings of performance on three dimensions (see APPENDIX C). The first two dimensions were selected from performance dimensions already used in the organization of interest. Thus, each organization used a different set of performance dimensions. These dimensions were selected in

conjunction with each organization's vice-president, personnel manager, and/or human resources manager. Due to the inclusion of the absolute comparison standard, only performance dimensions which were objectively quantified by the company were considered for selection. Furthermore, the company officers also assisted in determining the range of minimum, maximum, and/or average performance standards for each performance dimension which were used as goal levels for the absolute comparison standard and again for the multiple comparison standard. Some examples of actual organizational performance dimensions included: attendance, daily/monthly transaction rates, balancing record, hourly productivity, total sales, product knowledge, and selling cost. The performance dimensions varied widely with no dimension appearing more than twice (despite organizations of similar type) across the different companies. The third dimension was Overall Performance which was defined as the overall job performance when considering both of the previous dimensions. This dimension was included irrespective of whether or not the organization used it as a formal dimension in its performance appraisal process.

All of the dimensions were rated on a 9-point graphic rating scale (which allows for adequate variability in responses) anchored with 1 = Very Poor, 3 = Poor, 5 =

Average, 7 = Good, and 9 = Very Good. Each subject filled out five rating sheets. Each rating sheet used a different comparison standard as evidenced in the rating sheet instructions (see APPENDIX D). The comparison standard used and the instructions for the raters were as follows: (1) AMBIGUOUS - "Based on your performance over the past six months, please rate yourself on the following performance dimensions.", (2) INTERNAL - "Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your own personal, internal values and standards as a criteria. That is, base your ratings on how well you personally feel you have done over the past six months relative to your abilities and past performance. DO NOT give consideration to any other criteria beyond your own beliefs as to how well you performed.", (3) ABSOLUTE - "Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your company's minimum requirement or goal as the criterion. That is, for each dimension rate yourself in comparison to the minimal level of performance as defined by your company or group's policy. DO NOT give consideration to any other criteria beyond your own belief as to whether or not you met this minimum requirement.", (4) RELATIVE - "Based on your performance over the past six months,

please rate yourself on the following performance dimensions. Use your fellow coworkers' performance as a criterion. That is, think about how your co-workers have performed and compare yourself to them. DO NOT give consideration to any other criteria beyond your own belief as to how well you performed in direct comparison to your co-workers.", (5) MULTIPLE - "Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your own personal standards, your attainment of the minimum requirements and goals, and your comparison with fellow co-workers as the criteria. That is, consider all three standards as defined in the previous pages. Give equal consideration to all three of the criteria.

With the exception of the ambiguous rating sheet, all rating sheets were titled with their appropriate comparison standard. To control for order effects, the internal, relative, and absolute ratings sheets were presented in a randomized order. The ambiguous rating sheet was always presented first because it represented an undefined comparison standard, whereas the multiple rating sheet was always presented last because it represented a combination of the internal, relative, and absolute standards.

Supervisor evaluations. The supervisor rated the subordinate on the same five rating sheets and used the same three dimensions that the subordinate used to make self-ratings (including Overall Performance) with changes in the wording of the instructions appropriate for the supervisor (see APPENDIX E). The supervisory ratings were made on a graphic rating scale identical to the self-evaluations and were randomized similar to the self-evaluations.

Post-rating comparison standard ratings. After completing all of the ratings, each subordinate responded to the following question, "If asked to evaluate your own performance in the future, please rate each of the four comparison standards as to your preference for using them in future performance ratings". Supervisors responded to a similar questions, "If asked to rate employees in the future, please rate each of the four comparison standards as to your preference for using them in future performance appraisals". The question provided the Internal, Absolute, Relative, and Multiple comparison standards for the rater (see APPENDIX F).

Availability ratings. Each subordinate (i.e., self-rater) rated the availability of each comparison standard on a 5-point graphic rating scale with 1 = Not Available, 3 = Moderately Available, and 5 = Very Available (see

APPENDIX G). Availability was defined as the degree to which information pertinent to the comparison standard could be readily and easily obtained.

Relevancy ratings. Each subordinate (i.e., self-rater) rated the relevancy of each comparison standard on a 5-point graphic rating scale with 1 = Not Relevant, 3 = Moderately Relevant, and 5 = Very Relevant (see APPENDIX H). Relevancy was defined as the degree to which the comparison standard is appropriate and applicable within the workplace as a basis for performance ratings.

Demographics and comprehension question. Each rater was asked to provide the following personal information: (1) Age, (2) Sex, (3) Job Title/Occupation, (4) Tenure with Company, and (5) Tenure with Present Job. In addition, supervisors were asked to indicate the number of subordinates under their direct supervision (see APPENDIX I).

A final question was used to assess the rater's comprehension and honesty in understanding and using the rating packet. Raters were asked to respond in a yes or no fashion to the following question, "Do you feel you understood all the instructions and questions asked throughout this packet and were able to answer them in an honest and accurate manner?".

RESULTS

The means and standard deviations for the self-supervisor performance ratings are shown in Table 3. The data have been arranged to include the entire sample as well as self- and supervisory rating sources across the three performance dimensions and five comparison standards. The averaged rating (mean of the three performance dimensions) for each comparison standard was also included. In addition to providing an additional data point, the averaged ratings also allowed for easier comparisons across comparison standards especially in those instances when there was no significant main effect for the performance dimensions.

Preliminary Analyses

Prior to an examination of the five hypotheses, an initial analyses was conducted to examine the effects of the individual companies in relation to the performance dimension mean for each comparison standard across the two rating sources. Table 4 presents the cell means (performance dimension mean) for each of the comparison standards across the nine companies.

This preliminary analysis was investigated using a 2 x 5 x 9 (rater source x comparison standard x organization) repeated measures ANOVA. To safeguard against violations of sphericity and an inflated Type I error rate, the

Table 3

Means and Standard Deviations for Self-Supervisor Performance Ratings

RATING	SELF	SUPERVISOR	FULL SAMPLE
	AMBIGUOUS Standard		
DIMENSION 1	6.53 (1.83)	6.52 (1.97)	6.52 (1.90)
DIMENSION 2	6.40 (2.05)	6.60 (1.98)	6.50 (2.01)
DIMENSION 3	6.80 (1.55)	6.61 (1.68)	6.71 (1.61)
AVERAGE	6.58 (1.81)	6.58 (1.88)	6.58 (1.48)
	INTERNAL Standard		
DIMENSION 1	7.06 (1.61)	6.99 (1.46)	7.02 (1.53)
DIMENSION 2	6.79 (1.94)	6.95 (1.73)	6.87 (1.83)
DIMENSION 3	7.15 (1.47)	7.08 (1.32)	7.11 (1.40)
AVERAGE	7.00 (1.67)	7.01 (1.88)	7.00 (1.43)

(table con'd)

Table 3 (con'd)

RATING	SELF	SUPERVISOR	FULL SAMPLE
	ABSOLUTE Standard		
DIMENSION 1	6.84 (1.87)	6.64 (2.01)	6.74 (1.94)
DIMENSION 2	6.57 (2.05)	6.65 (1.99)	6.61 (2.02)
DIMENSION 3	6.95 (1.46)	6.59 (1.63)	6.77 (1.55)
AVERAGE	6.79 (1.80)	6.63 (1.87)	6.71 (1.45)
	RELATIVE Standard		
DIMENSION 1	7.03 (1.72)	6.67 (1.93)	6.85 (1.83)
DIMENSION 2	6.73 (1.82)	6.80 (1.87)	6.76 (1.85)
DIMENSION 3	7.16 (1.51)	6.76 (1.75)	6.96 (1.64)
AVERAGE	6.97 (1.68)	6.74 (1.85)	6.86 (1.55)

(table con'd)

Table 3 (con'd)

RATING	SELF	SUPERVISOR	FULL SAMPLE
	MULTIPLE	Standard	
DIMENSION 1	7.08 (1.55)	6.75 (1.57)	6.91 (1.57)
DIMENSION 2	6.79 (1.87)	6.82 (1.70)	6.81 (1.79)
DIMENSION 3	7.16 (1.33)	6.84 (1.49)	7.00 (1.42)
AVERAGE	7.01 (1.59)	6.80 (1.59)	6.91 (1.30)

Note. n = 106 for Self and Supervisor; N = 212 for Full Sample. Standard deviations in parentheses.

Table 4

Organization x Comparison Standard Cell Means for Averaged Performance Dimensions

ORGANIZATION	AMB	INT	ABS	REL	MULT
COMPANY 1	5.32	6.29	6.02	5.44	5.98
	(1.75)	(1.22)	(1.51)	(1.91)	(1.41)
COMPANY 2	6.30	6.77	6.33	6.33	6.44
	(1.65)	(1.44)	(1.64)	(1.74)	(1.51)
COMPANY 3	7.53	7.20	7.43	7.63	7.50
	(1.22)	(.93)	(1.20)	(1.31)	(1.00)
COMPANY 4	6.47	6.83	6.26	6.83	6.72
	(1.68)	(1.51)	(1.77)	(1.75)	(1.46)
COMPANY 5	6.71	7.38	6.84	6.78	7.14
	(1.87)	(1.51)	(1.82)	(1.94)	(1.63)
COMPANY 6	5.96	6.19	6.55	6.51	6.54
	(1.76)	(1.78)	(1.83)	(1.58)	(1.69)
COMPANY 7	7.42	7.65	7.37	7.72	7.57
	(1.40)	(.98)	(1.33)	(1.05)	(1.16)
COMPANY 8	7.09	7.47	6.55	7.50	7.33
	(1.68)	(1.33)	(1.77)	(1.42)	(1.40)
COMPANY 9	6.08	7.25	7.05	7.75	7.33
	(2.24)	(1.16)	(.83)	(1.06)	(1.18)

Note. AMB = Ambiguous; INT = Internal; ABS = Absolute; REL = Relative; MULT = Multiple. Standard deviations in parentheses.
N = 212.

Huynh-Feldt epsilon was used (when necessary) to adjust between-subject and error degrees of freedom values when computing the significance of the *F* ratio. The ANOVA results are reported in Table 5.

Of particular note is the significant two-way interaction between the individual organizations and comparison standards. Significant mean differences can be seen both across companies and across comparison standards as reflected in the significant main effects. However, there was no main effect for rater source, which eliminates leniency as a potential problem, nor were there any significant interactions involving rater source. The significant interaction, which is illustrated in Figure 1, highlights the fact that the companies as a whole significantly differed in their ratings across the performance dimensions (when expressed as an average) as a function of which comparison standard was being considered. However, no discernible trend between company and standard was evident in the interaction.

With regard to this preliminary analysis, it should be noted that the individual dimensions themselves (i.e., the three performance dimensions) are not of any real importance. Each of the nine companies used their own performance dimensions preventing any appropriate measures of comparison across the three dimensions. Rather, the

Table 5

Repeated Measures ANOVA for Organizational Effects on Performance Ratings: Rater Source x Comparison Standard x Organization

SOURCE	df	MS	F	F_{cv}	*p*
RATER Effect	1	5.71	.28	3.89	.60
ORGANIZATION Effect	8	100.43	4.95	1.98	.001
RATER x ORG.	8	12.69	.63	1.98	.76
ERROR	194	20.31	---	----	---
STANDARD Effect	4	13.65	7.94	2.39	.001
RATER x STANDARD	4	1.38	.80	2.39	.52
ORG. x STAND.	32	3.40	1.98	1.48	.001
RATER x STANDARD x ORGANIZATION	32	1.13	.66	1.48	.93
ERROR	776	1.72	----	----	---

Note. N = 212.

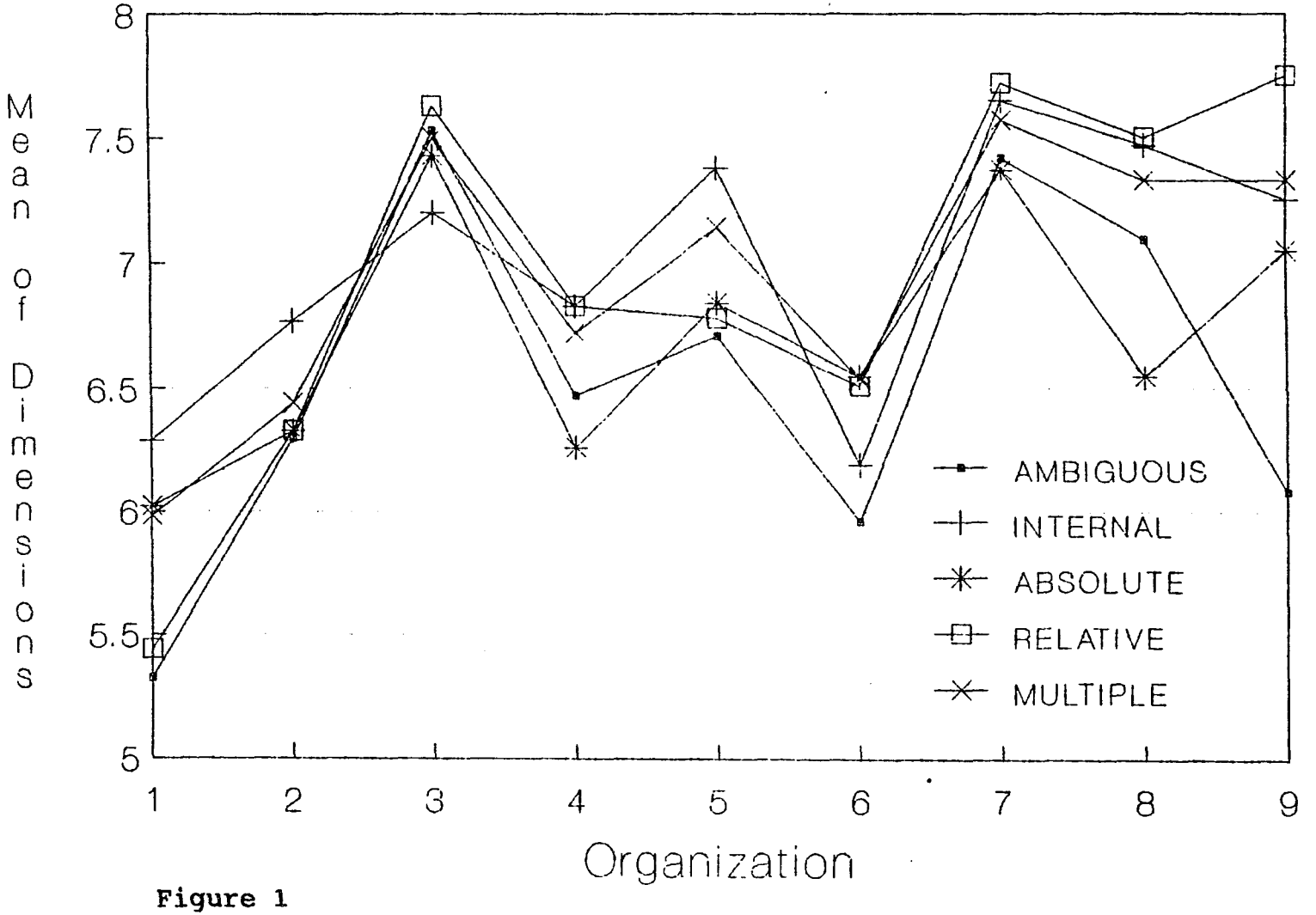

Figure 1

Organization x Comparison Standard Interaction

level of agreement between the self- and supervisory raters was the important issue with the individual performance dimensions being a means to an end. That is, the separate and distinct performance dimensions were used to pair the self-supervisor responses such that the ratings would be made on a dimension for the appropriate dyad within each company.

Hypothesis 1

The first hypothesis examined whether or not comparison standards were having a significant influence on performance ratings. It predicted that there would not be a main effect for rater source nor for the performance dimensions. However, a main effect for comparison standards was predicted indicating significant differences in rater responses, dependent on the comparison standard used. No significant interactions were predicted.

The hypothesis was tested by a 2 x 3 x 5 (rater source x performance dimensions x comparison standards) repeated measures ANOVA. The Huynh-Feldt epsilon was again used (when necessary) to safeguard against violations of sphericity and an inflated Type I error rate when analyzing the repeated measures ANOVA. The results of this ANOVA are presented in Table 6.

The main effect for the comparison standards directly confirms Hypothesis 1. Significant mean differences were

Table 6

Repeated Measures ANOVA for Performance Ratings: Rater Source x Performance Dimension x Comparison Standard

SOURCE	df	MS	F	F_{cv}	*p*
RATER Effect	1	11.11	.48	3.89	.49
ERROR	210	23.06	---	----	---
STANDARD Effect	4	18.20	10.35	2.50	.001
RATER x STANDARD	4	2.05	1.17	2.50	.32
ERROR	840	1.76	----	----	---
DIMENSION Effect	2	10.60	2.00	3.89	.14
RATER x DIMENSION	2	10.72	2.02	3.89	.13
ERROR	420	5.31	----	----	---
STANDARD x DIM.	8	.22	.39	2.10	.93
RATER x STAND. x DIM.	8	.23	.41	1.94	.92
ERROR	1680	.57	---	----	---

Note. N = 212.

detected across the five comparison standards. As anticipated, none of the variables produced significant interactions. A Tukey's HSD multiple comparison procedure was used to identify specific group mean differences in the comparison standards. The Tukey's critical difference (CD) value was adapted for repeated measures comparisons by replacing the MS_{WITHIN} with MS_{ERROR} and replacing *n* with N. Only the averaged ratings across the three performance dimensions were considered since no main effect for performance dimension was found. The results of this analysis are illustrated in Table 7.

The Tukey's findings suggest that ratings made on the ambiguous comparison standard are significantly lower than ratings taken from the internal comparison standard or the multiple comparison standard. Averaged ratings from the ambiguous comparison standard approached a significant mean difference when compared to the relative standard mean. The internal comparison standard produced the highest ratings followed by the multiple, relative, and absolute standards respectively; none of which significantly differed from one another.

The combined results of the repeated measures ANOVA and Tukey's HSD analyses are supportive of Hypothesis 1 and the associated predictions. No rater differences were detected across comparison standards indicating an absence

Table 7

Tukey's HSD Analysis of Comparison Standard Means

RELATIONSHIP TESTED	MEAN DIFFERENCE
AMBIGUOUS = INTERNAL	-.42**
AMBIGUOUS = MULTIPLE	-.33*
AMBIGUOUS = RELATIVE	-.28
ABSOLUTE = MULTIPLE	-.20
ABSOLUTE = RELATIVE	-.15
AMBIGUOUS = ABSOLUTE	-.13
RELATIVE = MULTIPLE	-.05
INTERNAL = MULTIPLE	.09
INTERNAL = RELATIVE	.14
INTERNAL = ABSOLUTE	.29

Note. N = 212.
* $p < .05$, Critical difference (CD) value = .33.
** $p < .01$, Critical difference (CD) value = .40.

of leniency on the part of self-raters. In fact, cursory examination of the means in Table 3 indicates that supervisor means were equal to or greater than self-rater means in two of the five comparison standards (40%) for averaged performance ratings.

Hypothesis 2

The second hypothesis examined the correlational relationships between the four explicit comparison standards, the ambiguous comparison standard, and previous self-supervisory relationships from the literature. It had been predicted that the four explicit comparison standards would yield significantly greater self-supervisor correlation coefficients than the ambiguous standard.

This hypothesis was investigated using the Pearson product-moment correlation coefficients (r) for self- and supervisory performance ratings across the various comparison standard formats. These correlations were tested against previous self-supervisor correlations in the literature as well as by direct comparison between the standards. Table 8 reports the correlations between the two rating sources for both the individual performance dimensions as well as the mean performance rating when collapsed across all three dimensions for each of the four explicit comparison standards and the ambiguous standard.

Table 8

Self-Supervisor Correlations Among Comparison Standards

MEAN	DIMENSION 1	DIMENSION 2	DIMENSION 3
AMBIGUOUS Standard			
.26	.31	.38	.26
INTERNAL Standard			
.43	.58	.45	.31
ABSOLUTE Standard			
.50	.63	.47	.49
RELATIVE Standard			
.43	.38	.42	.43
MULTIPLE Standard			
.55	.56	.53	.52

Note. All correlations were significant at the .01 significance level.
N = 212.

The first part of Hypothesis 2 was tested using a two-sample independent test for correlations ($r_1 = r_2$) comparing each explicit comparison standard to the ambiguous rating format. The internal comparison standard was found to produce a significantly greater self-supervisor correlation than the ambiguous standard ($z = 1.98$, $p < .05$) as did the absolute standard ($z = 2.89$, $p < .01$), the relative standard ($z = 1.98$), $p < .05$), and the multiple comparison standard ($z = 3.59$, $p < .001$) using one-tailed tests of significance. Thus, all four explicit standards produced higher interrater agreement which is highly supportive of Hypothesis 2a.

In addition, the absolute and multiple comparison standards (which did not significantly differ from one another) were found to produce significantly greater self-supervisor correlations than either the relative or internal standards (which did not significantly differ from one another.

The second part of Hypothesis 2 examined the relationship between the previous self-supervisor correlation ($r = .35$) in Harris and Schaubroeck's (1988) meta-analysis and the self-supervisor correlations for the four explicit comparison standards in this study. The four explicit comparison standards were expected to

produce significantly greater self-supervisor correlations than the meta-analysis coefficient.

Hypothesis 2b was tested by using a one-sample test for correlations with a constant (r_i = a) where a equalled .35. Both the absolute comparison standard ($z = 2.49$, $p < .01$), and the multiple comparison standard ($z = 3.49$, $p < .001$) generated significantly greater self-supervisor correlation coefficients than the literature constant of .35. However, neither the internal ($z = 1.20$, ns) nor the relative comparison standard ($z = 1.20$, ns) reached statistical significance when compared against Harris and Schaubroeck's meta-analysis findings. This finding is moderately supportive of the Hypothesis 2b, suggesting that the absolute and multiple comparison standards are particularly adept at increasing rater agreement on job performance while correlations generated from the internal and relative standard ratings were not significantly greater than .35.

The last part of Hypothesis 2 sought to establish that no statistically significant difference existed between the ambiguous comparison standard correlation coefficient and previous self-supervisor correlations in the literature. Again, a one-sample test for correlations with a constant (r_i = a) where a equalled .35 was used.

No statistically significant differences between the two correlation coefficients were expected.

This relationship was confirmed as the ambiguous standard ($z = 1.61$, ns) did not significantly differ from the literature's self-supervisor correlation of .35. Thus, the results are supportive of Hypothesis 2c.

Hypothesis 3

The third hypothesis examined the distribution of responses to the open-ended question concerning the basis of raters' current comparison standards. Hypothesis 3a predicted that self-raters would prefer an equivalent of the internal comparison standard while Hypothesis 3b predicted that supervisory raters would prefer an equivalent of the multiple comparison standard. A chi-square analysis was used to test the predictions. The expected and observed percentages of the chi-square tests are shown in Table 9.

Although the chi-square results were significant for both self-raters, $X^2(5, N = 106) = 65.5$, $p < .001$ and supervisory raters, $X^2(5, N = 106) = 70.0$, $p < .001$, the observed frequencies did not represent the expected pattern for either rater source. Instead, the "Other" category clearly dominated both distributions with the absolute standard being the preferred referent of the four explicit comparison standards for both rater types.

Table 9

Observed and Expected Frequencies for the Basis of Performance Ratings

COMPARISON STANDARD USED AS BASIS	f	EXPECTED %	OBSERVED %
	SELF-RATERS		
INTERNAL	9	17.7	8.5
ABSOLUTE	13	17.7	12.3
RELATIVE	3	17.7	2.8
MULTIPLE	11	17.7	10.4
OTHER	45	17.7	42.5
N/A	25	17.7	23.5
	n = 106		
	SUPERVISOR RATERS		
INTERNAL	4	17.7	3.8
ABSOLUTE	24	17.7	22.7
RELATIVE	7	17.7	6.6
MULTIPLE	8	17.7	7.5
OTHER	46	17.7	43.4
N/A	17	17.7	16.0
	n = 106		

The results were also unusual in regard to the high percentage of "N/A" responses for both supervisors and self-raters which indicated either a failure to answer the open-ended question or an inappropriate response.

An additional analysis was then performed to further investigate Hypothesis 3. In this second analysis, the final two categories (Other and N/A) were eliminated and the chi-square analysis was conducted on only the four explicit standards. This was done in hopes that with the elimination of unwanted categories, a more accurate interpretation of the results would be allowed. The exploratory chi-square results are presented in Table 10.

The chi-square results were significant for supervisory raters, $X^2(3, N = 43) = 22.6$, $p < .001$ but not for self-raters, $X^2(3, N = 36) = 6.2$, $p < .10$. While the order of standards remained unchanged with respect to the original chi-square findings, supervisor raters did prefer the absolute comparison standard significantly more when compared to the other explicit standards. For self-raters, there was no significant variation in the selection of rater bases for performance appraisal ratings. Combined, the *a priori* bases for rater comparisons illustrates a general reliance on the objective and goal-oriented absolute comparison standard.

Table 10

Observed and Expected Frequencies for the Basis of Performance Ratings (Condensed Version)

COMPARISON STANDARD USED AS BASIS	f	EXPECTED %	OBSERVED %
	SELF-RATERS		
INTERNAL	9	25.0	25.0
ABSOLUTE	13	25.0	36.1
RELATIVE	3	25.0	8.3
MULTIPLE	11	25.0	30.6
	n = 36		
	SUPERVISOR RATERS		
INTERNAL	4	25.0	9.3
ABSOLUTE	24	25.0	55.8
RELATIVE	7	25.0	16.3
MULTIPLE	8	25.0	18.6
	n = 43		

Hypothesis 4

The fourth hypothesis examined the preferences of both supervisors and subordinates in relation to future usage for each of the four explicit comparison standards. Hypothesis 4a predicted that both rating sources would prefer the multiple standard. Hypothesis 4b indicated that self-raters would next prefer the internal, absolute,and relative standards while Hypothesis 4c predicted supervisors would prefer absolute, relative, and internal standards after the multiple comparison standards. It had been anticipated that there would be a small, but significant main effect for rater source. A strong main effect for comparison standards was predicted indicating significant differences in rater preference across the four explicit comparison standards. Finally, a significant interaction had been hypothesized for the rater source and comparison standard relationship since mean differences in preferences would depend on rater source as well as the comparison standard in question. The means and standard deviations for the comparative standard preference ratings are shown in Table 11.

Hypothesis 4 was investigated using a 2 x 4 (rater source x comparison standard) repeated measures ANOVA to ascertain mean differences in preference ratings for potential use of the four explicit comparison standards in

Table 11

Means and Standard Deviations for Self-Supervisor Preference Ratings

COMPARISON STANDARD	SELF	SUPERVISOR	FULL SAMPLE
INTERNAL	6.28 (1.68)	5.59 (2.22)	5.93 (1.99)
ABSOLUTE	6.50 (1.44)	6.64 (1.85)	6.57 (1.66)
RELATIVE	6.26 (1.70)	5.93 (2.03)	6.09 (1.88)
MULTIPLE	7.03 (1.20)	6.76 (1.65)	6.90 (1.45)

Note. n = 106 for Self and Supervisor; N = 212 for Full Sample. Standard deviations in parentheses.

future performance appraisals. The Huynh-Feldt epsilon was again used (when necessary) to safeguard against violations of sphericity and an inflated Type I error rate when analyzing the repeated measures ANOVA. The results of this ANOVA are presented in Table 12.

The ANOVA findings indicated a significant interaction between rater source and the four explicit comparison standards. This interaction can be seen in Figure 2.

In general, self-raters gave higher preference ratings than supervisors except with the absolute comparison standard, although that difference was not significant at the $p < .05$ level. The main effect for comparison standards was statistically significant indicating mean differences in rater preferences for specific comparison standards.

The ANOVA was followed by a Tukey's HSD multiple comparison procedure. The Tukey's CD value was again adapted for repeated measures comparisons. These comparisons are reported in Table 13.

In general, the multiple standard appeared to be the preferred choice of raters. The multiple comparison standard approached statistical significance over the absolute standard when the two means were compared but fell short of the $p < .05$ cutoff. Nonetheless, in all cases, the multiple and absolute comparison standards

Table 12

Repeated Measures ANOVA for Preference Ratings: Rater Source x Comparison Standard

SOURCE	df	MS	F	F_{cv}	*p*
RATER Effect	1	17.55	3.13	3.89	.08
ERROR	210	5.61	---	----	---
STANDARD Effect	4	41.41	18.77	2.62	.001
RATER x STANDARD	4	6.27	2.84	2.38	.04
ERROR	630	1.76	----	----	---

Note. N = 212.

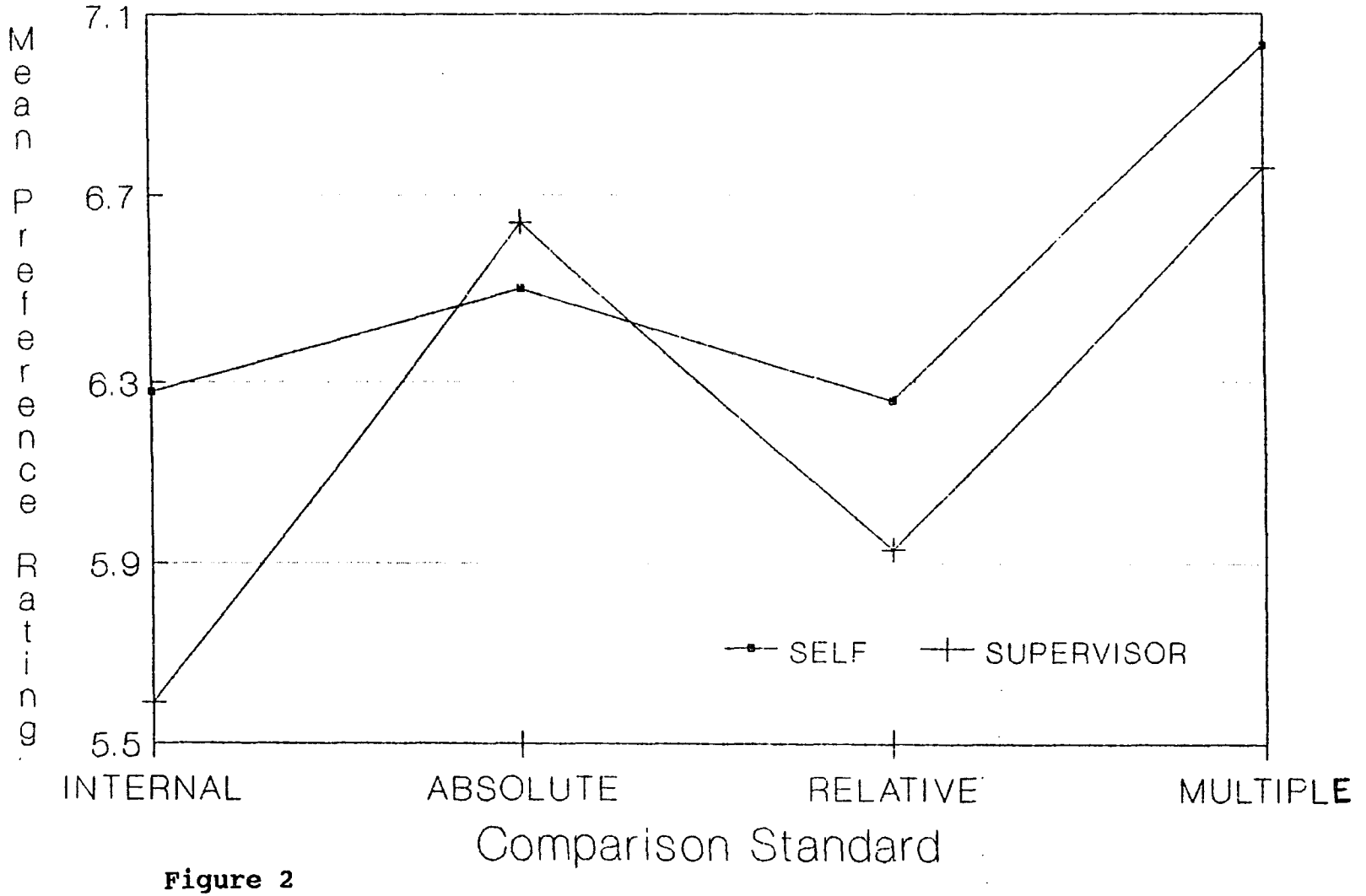

Figure 2

Rater x Comparison Standard Interaction

Table 13

Tukey's HSD Analysis of Preference Rating Means

RELATIONSHIP TESTED	MEAN DIFFERENCE (SUPERVISOR)	MEAN DIFFERENCE (SELF)	ABSOLUTE MEAN DIFFERENCE (FULL SAMPLE)
INTERNAL = MULTIPLE	-1.17**	-.75**	.97**
RELATIVE = MULTIPLE	- .83**	-.77**	.81**
INTERNAL = ABSOLUTE	-1.05**	-.22	.64**
ABSOLUTE = RELATIVE	.71**	.24	.48**
ABSOLUTE = MULTIPLE	- .12	-.53**	.33*
INTERNAL = RELATIVE	- .34*	.02	.16

Note. N = 212.
* $p < .05$, Critical difference (CD) = +/- .33.
** $p < .01$, Critical difference (CD) = +/- .40.

(which did not significantly differ in their preference means) were rated significantly higher in preference than the internal and relative standards (which did not significantly differ in their preference means) over the total sample. Furthermore, the multiple standard was preferred by almost one whole point on the 9-point rating scale over the internal standard for the full sample. Taken together, the combination of the ANOVA and Tukey's HSD findings were highly supportive of Hypothesis 4.

Hypothesis 4a was confirmed as both self-rater and supervisor subgroups indicated a preference towards using the multiple comparison standard in future job performance ratings. Hypothesis 4b received little support since, contrary to *a priori* predictions, self-raters preferred the absolute standard (after the multiple comparison standard) followed by the internal and relative standards respectively. Thus, Hypothesis 4b was not supported in the sense that the absolute and internal standards were reversed from their predicted order. Hypothesis 4c received strong empirical support as supervisory rating preferences exactly mirrored the predicted order (multiple, absolute, relative, and internal) for mean differences.

Overall, the fourth hypothesis indicated a preference for both self-raters and supervisors towards the multiple

standard and absolute comparison standards. Furthermore, the presence of the significant interaction between rater source and comparison standards was genuinely supportive of all three sub-hypotheses which suggested that rater preferences were substantially influenced by rater source and, more importantly, comparison standards.

Hypothesis 5

The fifth and final hypothesis explored the effects of availability and relevancy and their effects on rater source, comparison standards, and preference ratings. For ease of interpretation, availability and relevancy will be referred to as "referent dimensions" when discussed as a single factor.

No significant main effect was predicted for rater source as neither availability nor relevancy ratings were expected to fluctuate as a result of this variable. However, strong main effects for both comparison standards and the two referent dimensions were expected since the theoretical work of Kulik and Ambrose (1992) suggested such a phenomenon. Of particular interest was the interaction between the two referent dimensions (availability and relevancy) and the comparison standards. It was predicted that this interaction would be significant. The three interactions involving rater source were predicted to be significant because of the

powerful effects from the referent dimensions as well as the effects from the comparison standards. Table 14 displays the means and standard deviations of the availability and relevancy ratings across rater source and comparison standards.

Hypothesis 5 was investigated by generating a 2 x 2 x 4 (rater source x referent dimension ratings x comparison standards) repeated measures ANOVA. The ultimate purpose of this analysis was to identify whether a relationship existed between the rater's preference for each comparison standard and the referent dimensions of that standard. The within-subjects ANOVA was safeguarded by the Huynh-Feldt epsilon for violations of sphericity when appropriate. The results of the ANOVA are presented in Table 15.

The ANOVA findings indicated a significant three-way interaction for rater source, referent dimensions, and comparison standards. An examination of this significant interaction in Figure 3 indicates that relevancy ratings from supervisors were low on the internal standard and high on the absolute standard as were supervisory availability ratings. Furthermore, supervisory ratings were higher than self-ratings in all instances except the internal standard for both availability and relevancy. Thus, the availability and relevancy ratings are dependent

Table 14

Means and Standard Deviations for Self-Supervisor Availability and Relevancy Ratings

COMPARISON STANDARD	SELF	SUPERVISOR	FULL SAMPLE
	AVAILABILITY		
INTERNAL	4.18 (1.07)	3.92 (1.12)	4.05 (1.10)
ABSOLUTE	4.17 (1.06)	4.44 (.82)	4.31 (.95)
RELATIVE	3.82 (1.15)	4.02 (.95)	3.92 (1.05)
MULTIPLE	4.06 (.83)	4.06 (.75)	4.06 (.79)
	RELEVANCY		
INTERNAL	3.92 (1.00)	3.36 (1.30)	3.64 (1.19)
ABSOLUTE	4.11 (.93)	4.31 (.76)	4.21 (.86)
RELATIVE	3.49 (1.17)	3.54 (1.04)	3.51 (1.11)
MULTIPLE	3.94 (.84)	3.96 (.84)	3.95 (.84)

Note. n = 106 for Self and Supervisor; N = 212 for Full Sample. Standard deviations in parentheses.

Table 15

Repeated Measures ANOVA for Availability and Relevancy Ratings: Rater Source x Referent Dimensions x Comparison Standard

SOURCE	df	MS	F	F_{cv}	*p*
RATER Effect	1	.05	.02	3.89	.89
ERROR	210	2.49	---	----	---
STANDARD Effect	3	14.17	16.67	2.62	.001
RATER x STANDARD	3	1.50	1.76	2.62	.15
ERROR	630	.85	----	----	---
REF. DIM. Effect	1	52.68	49.23	3.56	.001
REF. DIM. x RATER	1	7.53	7.04	3.56	.01
ERROR	210	1.07	----	----	---
STANDARD x REF. DIM.	3	3.97	6.84	2.61	.001
RATER x STANDARD x REF. DIM.	3	5.43	9.31	2.61	.001
ERROR	630	.58	---	----	---

Note. REF. DIM. = Referent dimensions (availability and relevancy ratings).
N = 212.

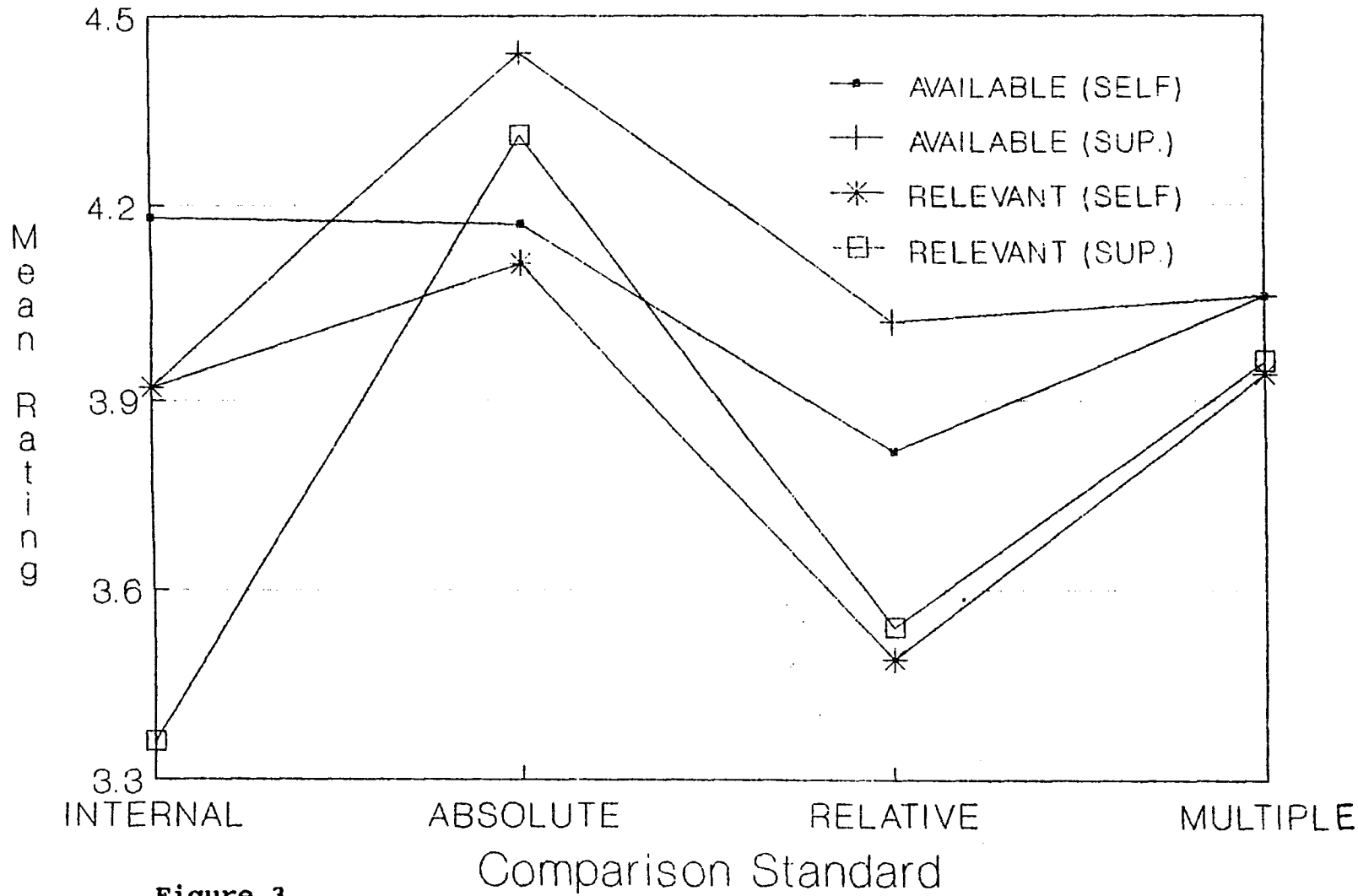

Figure 3

Rater Source x Referent Dimensions x Comparison Standard Three-Way Interaction

on the source of the rater and the specific comparison standard. Such a finding is very favorable in regard to the confirming the hypothesis that differential effects are occurring.

No significant effect for rater source was found. However, the main effect for comparison standards and the main effect for referent dimensions were both statistically significant. The main effects indicated that each of the comparison standards tended to be rated higher in availability than on relevancy. Furthermore, in all cases, the absolute comparison standard was rated highest in terms of availability and relevancy followed by the multiple, internal, and relevant standards respectively.

A Tukey's HSD post hoc comparison was then performed on the availability and relevancy variables so as to examine mean differences in the interaction cells, with the critical difference value converted for repeated measures comparisons. The results of these comparisons are reported in Table 16.

The absolute comparison standard was significantly more available than the other three standards. The absolute and multiple standards were each considered significantly more relevant than the other two comparison standards. However, the absolute comparison standard

Table 16

Tukey's HSD Analysis of Availability and Relevancy Means

RELATIONSHIP TESTED	MEAN DIFFERENCE (SUPERVISOR)	MEAN DIFFERENCE (SELF)	ABSOLUTE MEAN DIFFERENCE (FULL SAMPLE)
	AVAILABILITY		
ABSOLUTE = RELATIVE	.42**	.35**	.39**
INTERNAL = ABSOLUTE	-.52**	.01	.26*
ABSOLUTE = MULTIPLE	.38**	.11	.25*
RELATIVE = MULTIPLE	-.04	-.24*	.14
INTERNAL = RELATIVE	-.10*	.36**	.13
INTERNAL = MULTIPLE	-.14	.12	.01
	RELEVANCY		
ABSOLUTE = RELATIVE	.77**	.62**	.70**
INTERNAL = ABSOLUTE	-.95**	-.19	.57**
RELATIVE = MULTIPLE	-.42**	-.45**	.44**
INTERNAL = MULTIPLE	-.60**	-.02	.31**
ABSOLUTE = MULTIPLE	.35**	.17	.26*
INTERNAL = RELATIVE	-.18	.43**	.13

Note. N = 212.
* $p < .05$, Critical difference (CD) = +/- .23.
** $p < .01$, Critical difference (CD) = +/- .28.

emerged as the predominant favorite. Supervisory ratings appeared to differ significantly more across the standards as they produced more significant mean differences than the self-raters.

When viewed in conjunction with the results of the fourth hypothesis, there is moderate to strong support for Hypothesis 5 which indicates significant differences in comparison standard preferences based on availability and relevancy. However, the order of preference for the absolute and multiple comparison standards was reversed in relationship to their availability and relevancy ratings.

Supplemental Analysis (Hypothesis 5)

To better understand the nature of the relationship between raters' preferences and the referent dimensions of availability and relevancy, a multiple regression analysis was conducted to determine the effect of comparison standards, availability, relevancy, and their subsequent interactions on rater preferences.

A regression equation was generated using preference ratings as the dependent variable. Comparison standard categories were effect-coded and loaded into the equation as three dummy variables for the first step in the blocked regression. The second step consisted of the availability ratings and relevancy ratings. The third step represented the interactions between the comparison standards and

availability ratings. The interactions between comparison standards and relevancy ratings comprised the fourth and final step in the equation. The results of this hierarchial regression are reported in Table 17.

The results indicate that the referent dimensions (availability and relevancy) were accounting for the largest portion of the variance in rater preferences for use of comparison standards. The entire equation accounted for 32% of the total variability in preference ratings with a total R equal to .57. The two referent dimensions were responsible for almost 27% of the total variability alone. The comparison standards accounted for a small, but significant portion of the variance, while the interaction between the four explicit comparison standards and the referent dimension of availability added small, but significant, incremental variance. The interaction between relevancy and comparison standards did not produce any significant incremental variance.

This supplemental analysis lends additional credence to the contention that availability and relevancy factors figure prominently in raters' preferences for one comparison standard over another. This finding provides direct empirical support for the propositions of Kulik and Ambrose (1992).

Table 17

Multiple Regression Analysis: Effects of Comparison Standards, Availability, and Relevancy on Preference Ratings

Block 1 R^2	.05***	Block 1 ΔR^2	.05***
Comparison Standard			
D1	0.23***		
D2	- 0.08		
D3	- 0.19***		
Block 2 R^2	.31***	Block 2 ΔR^2	.26***
Referent Dimensions			
AVAILABILITY	0.19***		
RELEVANCY	0.42***		
Block 3 R^2	.32*	Block 3 ΔR^2	.01*
Availability Interactions			
D1 x AVAILABILITY	0.43**		
D2 x AVAILABILITY	0.11		
D3 x AVAILABILITY	0.27		
Block 4 R^2	.32	Block 4 ΔR^2	.00
Relevancy Interactions			
D1 x RELEVANCY	0.08		
D2 x RELEVANCY	- 0.08		
D3 x RELEVANCY	0.03		

Note. Standardized beta weights are reported in the table. D1-D3 represent dummy variables for comparison standards.
N = 848.
* $p < .05$. ** $p < .01$. *** = $p < .001$.

In conclusion, the overall trend of results is highly supportive, demonstrating the powerful impact differential comparison standards can have on self-supervisor performance appraisal ratings, especially when viewed in terms of preference, availability and relevancy, and significantly increased interrater agreement.

Exploratory Analysis

One additional analysis was performed in an effort to explore and provide additional information for one particular area of consideration.

This exploratory analysis examined the amount of observed halo in each of the five comparison standards. The mean correlation for each comparison standard was calculated using the intercorrelations between the three performance dimensions for both self- and supervisory raters. The purpose of this analysis was to determine whether any particular trends existed in regard to observed halo. The results are reported in Table 18.

These correlations suggest a strong relationship among the three performance dimensions in general, as reflected in the high mean intercorrelations (i.e., dimension similarity). However, it should be noted again that any interpretation of the correlations is tenuous due to the varying performance dimensions used across companies.

Table 18

Self- and Supervisory Mean Performance Dimension Intercorrelations across Comparison Standards

	SELF	SUPERVISOR
AMBIGUOUS	.74	.70
INTERNAL	.68	.64
ABSOLUTE	.71	.69
RELATIVE	.73	.69
MULTIPLE	.82	.76

Note. All correlations are significant at $p < .01$.

Furthermore, since no true scores are available, observed halo can only be examined in a relative sense.

One notable pattern was in the form of self-raters who consistently produced higher levels of observed halo than supervisors although the differences were not significant in any comparison. Additionally, the multiple standard generated higher observed halo correlations than the other four comparison standards. This is not surprising since the multiple comparison standard represented a composite of the other three explicit comparison standards. When the multiple standard is excluded, the ambiguous standard produced the highest set of intercorrelations (mean $\underline{r}$ = .72) suggesting that this standard is more likely to be prone to halo error.

DISCUSSION

The purpose of this research was to examine the differential effects of comparison standards and their impact on self- and supervisory ratings in a performance appraisal context. In general, it appears that the comparison standards produced both significant main effects and significant interactions in conjunction with a variety of other variables across several supported hypotheses. Furthermore, specific trends in rater preferences were distinctly discernible in terms of comparison standards as well as the availability and relevancy variables previously posited by Kulik and Ambrose (1992). A particularly pronounced finding involved the significantly increased interrater agreement when explicit comparison standards were compared to ambiguous comparison standards and the previous meta-analytic correlation coefficient reported by Harris and Schaubroeck (1988). These findings represent a significant and supportive step in identifying differential comparison standards as an underlying mechanism responsible for the poor interrater agreement which has typified self-ratings studies. Only the open-ended questions which explored raters' bases for referent selection produced less than supportive evidence for the impact of differential comparison standards.

Nevertheless, the net results, across all five hypotheses, strongly supported the existence of comparison standards and the significant influence they have on the agreement between self- and supervisory raters.

Interpretation of the Results

The findings from Hypothesis 1 indicate that raters are in fact rating performance differently dependent on which comparison standard is being considered. All four of the explicit comparison standards produced higher performance ratings on average than the ambiguous standard condition. Furthermore, both the internal and multiple standards produced significantly increased performance ratings above and beyond the ambiguous standard. While higher performance ratings, in and of themselves, do not suggest a better performance appraisal system, these results do suggest that the more vague and global instructions indicative of the ambiguous standard resulted in ratings which were more "average" (i.e., central tendency error). On the other hand, the explicit comparison standards, which relied on specific referent groups and more defined criteria, produced higher and more definitive assessments of performance behaviors.

The lack of any significant mean differences between self- and supervisory performance ratings across the standards further suggests that leniency was not an issue

nor was it responsible for increases in the explicit comparison standards' ratings. This finding lends additional credence to the earlier works of Farh and Werbel (1986) and Somers and Birnbaum (1991) which found self-ratings to be free of any significant leniency error. Again, the explicit and definitive nature of the instructions for each comparison standard may have improved the relationship between performance ratings for both rating sources.

The second hypothesis produced potentially more powerful results in terms of supporting the effects of comparison standards on self-supervisor rating agreement. Examination of all three sub-hypotheses collectively illustrates the primary advantage for the inclusion of comparison standards as a beneficial component of the performance appraisal system. All four explicit comparison standards (internal, absolute, relative, and multiple), when averaged across the three performance dimensions, produced significantly greater interrater agreement correlations than the ambiguous standard. In addition, all four explicit comparison standards produced correlation coefficients greater than the previous meta-analytic correlation produced by Harris and Schaubroeck (1988), although only the absolute and multiple standards were significantly greater. Furthermore, the ambiguous

standard did not significantly differ from Harris and Schaubroeck's correlation indicating that the ambiguous standard used in this study was potentially reflective of previous studies' operationalization of the rating instructions.

Clearly, the use of explicit comparison standards, which specifically define the referent group of interest, results in increased interrater agreement (ranging from .43 to .55) between self- and supervisory ratings in a performance appraisal context. Particularly, the absolute and multiple comparison standards, which both incorporate a specific goal level, represent the strongest comparison standards with correlations between self- and supervisory ratings equal to .50 and .55 respectively. It is likely, that by further defining the referent group of interest to include specific and attainable goals (cf., Locke's goal-setting paradigm; Latham & Locke, 1991) as in the absolute and multiple standards, the resulting product is a significant increase in the agreement between rating sources. Thus, even though researchers have long advocated qualitative differences in self- and supervisory schemas of performance (e.g., Bernardin & Beatty, 1984; McEnery & McEnery, 1987), it would appear that explicit comparison standards are capable of compensating for these differences by effectively producing similar reference

groups. Similar views have been advocated by McDonald (1990) who found that when self-raters were provided with similar referent group data and had knowledge of the performance dimensions beforehand, the result was increased interrater agreement and accuracy. Comparison standards in this sense then, are not unlike the frame-of-reference training system which also seek to bring rater viewpoints into harmony.

Hypothesis 3 produced the study's only unsupportive hypothesis. The "Other" category dominated the listing for both self- and supervisory raters. However, this may have been a function of the open-ended question format which allowed raters to select their own bases for making performance ratings prior to any exposure of known referent groups. Previous research often dictated a series of referent groups which raters could pick from (e.g., Stepina & Perrewe, 1991). A second possible influence is the operationalization of the "Other" category which included subjective assessments of personality traits either alone OR in combination with other comparison standards. Had these subjective assessments been operationalized to be included in the internal category (i.e., self) or allowed to be absorbed into the predominant comparison standard, in which they occurred in combination, the "Other" category would have

been significantly smaller and a completely different set of chi-square observed frequencies would have been produced altogether. Furthermore, just because raters like particular traits, that does not legitimize their use, especially when these subjective assessments run counter to good business practice.

Thus, the questioning format used (open-ended vs. multiple choice), specific knowledge of available referent groups (*a priori* vs. no previous knowledge), and operationalization of the referent group categories should all be important considerations in future comparison standard studies.

Results from the fourth hypothesis displayed a definite trend in rater preferences for comparison standard applications in performance appraisal systems. The previous work of Kulik and Ambrose (1992) and Oldham et al. (1986) was instrumental in predicting the rater preference patterns. The fact that both self-raters and supervisors preferred the absolute and multiple comparison standards (which were statistically equivalent) over the other comparison standards is particularly relevant in conjunction with Hypothesis 2 which found that the absolute and multiple comparison standards produced the greatest interrater agreement. Thus, not only were these two comparison standards the most preferred by both rating

sources they also produced the highest correlation coefficients in terms of rater agreement.

Additionally, the findings indicate that both self- and supervisor raters prefer the comparison standard which utilizes the most information. Because the multiple standard is comprised of the internal, relative, and absolute comparison standards, it represents the most comprehensive and informative referent group available. Apparently, not only did raters cue in on this, they also like the idea of increased reference points in making performance appraisal determinations. It is particularly interesting to note that the self-raters (i.e., subordinates) preferred the multiple and absolute comparison standards in parallel fashion to the supervisors. This suggests that the self-raters were more than willing to have specific and objective behavioral goals included in their performance appraisals. While such a finding is not surprising with regard to the supervisors, who must often rely on numerous objective criteria for decision-making, previous research (e.g., Heneman, 1986; Mabe & West, 1982) would suggest that subordinates often shy away from such definitive measures of their performance. Apparently, reliance on more subjective appraisal systems such as those incorporated

into the internal and relative comparison standards, are not as desirable to either the ratee or the rater.

Findings from the fifth hypothesis confirmed the previously theoretical propositions of Kulik and Ambrose (1992) who suggested that raters would prefer comparison standards which were readily available and relevant to the rater's situation. It was first established that the availability and relevancy variables were having an influence on the comparison standards. The significant main effect for comparison standards suggested that raters were, in fact, discriminating in terms of the availability and relevancy of the particular standard. Furthermore, across all standards, the availability and relevancy variables were being distinguished as separate constructs as evidenced by the significant main effect for availability and relevancy variables. Finally, the multiple regression analysis confirmed the fact that the referent dimensions of relevancy and availability were accounting for significant variability in rater preferences amongst comparison standards.

The significant three-way interaction of rater source, comparison standards, and availability/relevancy variables is of the most importance here. This interaction suggests a definite relationship between the availability and relevancy of a particular comparison standard in direct

comparison with other standards as well as a dependence on the rating source. The cell means indicate that overall, the absolute comparison standard was the most available and the most relevant, followed closely by the multiple comparison standard.

When Hypotheses 4 and 5 are considered together, it appears that the comparison standards which are the most preferred by raters also happen to be the comparison standards which are the most available and relevant.

Implications and Conclusions

This study has generated substantial theoretical and empirical support for the existence, classification, and effectiveness of differential comparison standards within a performance appraisal framework. Additionally, the findings underscored significant implications as to the how's and why's of raters' referent group selection processes of comparison standards by exploring pre-rating bases for referent choice and post-rating preferences. While previous theory and research suggested that internal and/or self-referents were the preferred comparative referent (e.g., Oldham et al., 1986; Summers & DeNisi, 1990), this study found support for raters preferring to use comparison standards which were more comprehensive, more objective, and goal-driven. In addition, raters also considered whether or not the basis for rating performance

was available and relevant to the evaluation at hand. In both instances, the evidence pointed to the absolute and multiple comparison standards representing the raters' preferred choice. The observed percentage results from the chi-square analyses illustrated the raters' preference even prior to their knowledge of the four explicit standards.

Furthermore, the study established a strong relationship between rater preferences and the importance of availability and relevancy in comparison standards which, in turn, provided empirical support for the Kulik and Ambrose (1992) and Oldham et al. (1986) proposition that relevancy and availability are important referent dimensions in determining how raters select comparison standards. Thus, equity theory, social comparison theory, and relative deprivation theory were all indirectly supported as they served as the basic foundation for Kulik and Ambrose's propositions.

Finally, it was shown that by adopting explicit comparison standards, correlation coefficients for self-supervisor interrater agreement greater than .35 are attainable and can even reach as high as .55. This finding counters the earlier works of Harris and Schaubroeck (1988) and Mabe and West (1982) and suggests that when performance ratings by multiple sources are made

under explicit comparison instructions on dimensions which are objectively quantifiable, the result is a significant enhancement of interrater agreement.

Examination of the five hypotheses in tandem suggests that adopting the absolute and/or multiple comparison standard as an integral part of the performance appraisal process would be highly advantageous for a variety of reasons. The second hypothesis established that the absolute and multiple comparison standard formats produced the greatest interrater agreement between self-raters and supervisors. This level of agreement ($\underline{r}$ >= .50) was substantially higher than previous studies in the literature as well as the non-explicit comparison standard. The fourth hypothesis identified the absolute and multiple comparisons standards as the preferred standards for both self-raters and supervisors. Finally, the fifth hypothesis discovered that both self- and supervisory raters identified the absolute and multiple comparison standards as the most available and relevant comparative choices. Adoption of these two comparison standards would likely be endorsed by all the literature which advocates inclusion of the self-rater in the performance appraisal process (e.g., Fletcher, 1986; Latham & Wexley, 1981; Riggio & Cole, 1992). Hence, the absolute and multiple comparison standards seem to

represent the best choice in all pertinent areas for use in performance appraisal systems.

Limitations

Despite some very interesting and provocative findings, there exist certain limitations within the study which may affect the reliability and validity of the results.

The first potential problem centers around the comprehension of the material by the subjects. That is, do they understand what is being asked of them. Results from the open-ended question asking for rater bases' of performance produced relatively high rates of non-applicable responses and/or no response to the question for both supervisors (16%) and self-raters (23.5%).

However, in defense of the first potential limitation, only 15 subjects in the total sample of 212 (7%) gave a "No" response to the final question which asked whether subjects had understood all the instructions/questions and accurately responded to them.

The second limitation revolves around the sample itself. The organizational sample was primarily comprised of banking institutions which tend to use performance dimensions which are very specific, very objective, and goal-driven. Similarly, all the companies included in the study represented organizations which already used the

operational equivalent of absolute comparison standards (performance was evaluated in relation to an established and objective goal) in their performance appraisal process. Thus, companies which relied more heavily on subjective appraisals or which had no formal appraisal system at all were automatically excluded. The cumulative result may have been a bias towards the absolute and multiple comparison standards which were already in effect within the companies. It is possible, then, that the powerful results supporting adoption of absolute and multiple standards (Hypothesis 2 - increased interrater agreement; Hypothesis 4 - preferred standard; and Hypothesis 5 - more available and relevant standards) are actually an artifactual result of the organizations employed in the study. Furthermore, organizations and/or job types which do not use or allow for objective criteria in regard to job performance many find it difficult to benefit from the advantages of explicit comparison standards outlined in this paper.

On the other hand, it may be that these companies have adopted absolute (and multiple standards) precisely because they are better standards of performance. Industrial/organizational psychology has always pushed for more objective, specific, and goal-driven measures of performance (Landy, 1989; Latham & Locke, 1991). The net

effect being that over time more organizations are using objective, goal-specific measures of performance which result in higher levels of agreement, preference, availability, and relevancy among raters. In essence, studies should expect a natural increase in self-supervisor interrater agreement as performance appraisal systems become more accurate, sophisticated, and objective.

Applications and Future Studies

It is hoped that this study has shown how the incorporation of differential comparison standards can be used in future studies to increase interrater agreement between various rating sources above and beyond previous research endeavors.

Researchers will have to consider the practical applications and effects of rating instructions in conjunction with differential comparison standards when designing or studying performance appraisal formats. Rating instructions which fail to distinguish the referent group of interest must be changed so as to incorporate the advantages of more explicit and objective criteria. Performance appraisal designers will need to redesign performance rating sheets to incorporate more explicit instructions which specifically indicate the comparison standard to be used. This must be done for all rating

sources. Based on the results from this study, the performance appraisal process would benefit significantly from the adoption of the absolute and/or multiple comparison standard replete with instructions which include goal levels for objective criteria. Future studies may wish to examine how peer raters respond to the absolute and multiple comparison standards. Encouraging findings would allow three different rating sources to more accurately evaluate job performance.

The shared importance of differential comparison standards and differential reference points will have to be addressed in future studies especially those involving multiple raters. The basic tenets of comparison standard research essentially model those of FOR training but from different ends of the same continuum. Later studies may wish to explore the commonalities and/or differential effects between comparison standards and reference points. This might be accomplished by employing frame-of-reference training techniques to comparison standard rater training. Of particular interest would be an empirical research design which assimilated both FOR training and comparison standard rating formats within the same study. Ideally such a combination could boost interrater agreement well past correlations of .55. Obviously, this would also advocate the use of rater training for supervisors,

subordinates, and peers in reference to a job-related context. Such a dynamic system would embrace all the advantages of both traditional and alternative performance appraisal systems including enhanced convergent validity.

Additional studies will be needed to determine how and why raters select the comparison standards they do beyond just availability and relevancy factors. The findings from Hypothesis 3 suggest that without prior conscious knowledge of explicit comparison standards, raters are using a basis of performance which does not model any single standard but reflects a combination of various factors. Enhanced operationalization of comparison standard classification schemes needs to be addressed (i.e., what to do with subjective assessments of performance and various combinations of objective and subjective bases of performance) if related studies are to be used in meta-analyses and/or more headway is to be made in determining exactly which referents are used by raters and how they are selected.

Finally, it is hoped that these findings will result in a renewed interest in the self-ratings area after several previous studies (e.g., Harris & Schaubroeck, 1988) have downplayed the importance and validity of self-ratings in performance appraisal research. Continued advancements in the laboratory and the business community

over the years is leading to more accurate and objective performance appraisal systems (Landy, 1989). With the continued revisions and improvements in these systems, self-ratings may yet return to play a pivotal role.

REFERENCES

Adams, J. S. (1963). Towards an understanding of inequity. Journal of Abnormal and Social Psychology, 67, 422-436.

Adams, J. S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), Advances in experimental social psychology, (Vol. 2, pp. 267-299). New York: Academic Press.

Arnold, J., & Davey, K. M. (1992). Self-ratings and supervisor ratings of graduate employees' competence during early career. Journal of Occupational and Organizational Psychology, 65, 235-250.

Ashford, S. J. (1989). Self-assessments in organizations: A literature review and integrative model. In L. L. Cummings & B. M. Staw (Eds.), Research in organizational behavior, 11, 133-174.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels of processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72, 567-572.

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. Journal of Applied Psychology, 77, 975-985.

Bassett, G. A., & Meyer, H. H. (1968). Performance appraisal based on self-review. Personnel Psychology, 21, 421-430.

Bernardin, H. J., & Beatty, R. W. (1984). Performance appraisal: Assessing human behavior at work. Boston, MA: Kent Publishing Company.

Bernardin, H. J., & Villanova, P. (1986). Performance appraisal. In E. A. Locke (Ed.), Generalizing from laboratory to field settings. Lexington, MA: Lexington Books.

Borman, W. C. (1974). The ratings of individuals in organizations: An alternative approach. Organizational Behavior and Human Performance, 12, 105-124.

Campbell, D. J., & Lee, C. (1988). Self-appraisal in performance evaluations: Development versus evaluation. Academy of Management Review, 13(2), 302-314.

Carroll, S. J., & Schneier, C. E. (1982). Performance appraisal and review systems: The identification, measurement, and development of performance in organizations. Glenview, IL: Scott, Foresman.

Cascio, W. F. (1987). Applied psychology in personnel management (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. Journal of Applied Psychology, 74, 130-135.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton Mifflin Co.

DeNisi, A. S., Cafferty, T., & Meglino, B. (1984). A cognitive view of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

DeNisi, A. S., & Williams, K. J. (1988). Cognitive approaches to performance appraisal. In G. Ferris and K. Rowland (Eds.), Research in personnel and human resource management, Vol. 6, Greenwich, CT: JAI Press.

Dornstein, M. (1989). The fairness judgments of received pay and their determinants. Journal of Occupational Psychology, 62, 287-299.

Farh, J., & Dobbins, G. H. (1989). Effects of comparative performance information on the accuracy of self-ratings and agreement between self- and supervisory ratings. Journal of Applied Psychology, 74(4), 606-610.

Farh, J., & Werbel, J. D. (1986). Effects of purpose of the appraisal and expectation of validation on self-appraisal leniency. Journal of Applied Psychology, 71(3), 527-529.

Farh, J., Werbel, J. D., & Bedeian, A. G. (1988). An empirical investigation of self-appraised based performance evaluation. Personnel Psychology, 41, 141-156.

Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.

Fisher, C. D. (1989). Self and superior assessment: Unraveling the causes of disagreement. Unpublished manuscript. University of Baltimore, Baltimore, MD.

Fletcher, C. (1986). The effects of performance review in appraisal: Evidence and implications. Journal of Management Development, 5(3), 3-12.

Ford, J. K., & Noe, R. A. (1987). Self-assessed training needs: The effects of attitudes toward training, managerial level, and function. Personnel Psychology, 39-53.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. Personnel Psychology, 41, 581-592.

Gioia, D. A., & Sims, H. P. (1986). Cognition-behavior connections: Attribution and verbal behavior in leader-subordinate interactions. Organizational Behavior and Human Decision Processes, 37, 197-229.

Goodman, P. S. (1974). An examination of the referents used in the evaluation of pay. Organizational Behavior and Human Performance, 12, 170-195.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisory ratings. Personnel Psychology, 41, 43-62.

Hauenstein, N. M., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. Personnel Psychology, 42, 359-378.

Henderson, R. I. (1984). Performance appraisal. Reston, VA: Reston Publishing Company.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. Personnel Psychology, 39, 811-826.

Heneman, R. L., Wexley, K. N., & Moore, M. L. (1988). Performance rating accuracy: A critical review. Journal of Business Research, 15, 431-448.

Hoffman, C. C., Nathan, B. R., & Holden, L. M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. Personnel Psychology, 44, 601-618.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. Psychological Bulletin, 96, 72-98.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw & L. L. Cummings (Eds.), Research in organizational behavior (Vol. 5, pp. 141-196). Greenwich, CT: JAI Press.

Jones, M. (1991). Self-assessment and supervisor assessment: A review. Unpublished manuscript. Louisiana State University, Baton Rouge, LA.

Jones, E. E., & Nisbett, R. E. (1971). The actor and the observer: Divergent perceptions of the causes of behavior. Morristown, N.J.: General Learning Press.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. Jones, D. Kanouse, H. Kelly, R. Nisbett, S. Valines, B. Weiner (Eds.), Attribution: Perceiving the causes of behavior (pp. 79-94). Morristown, NJ: General Learning Press.

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59(4), 445-451.

Kruglanski, A. W., & Mayseless, O. (1990). Classic and current social comparison research: Expanding the perspective. Psychological Bulletin, 108, 195-208.

Kulik, C. T., & Ambrose, M. L. (1992). Personal and situational determinants of referent choice. Academy of Management Review, 17, 212-237.

Landy, F. J. (1989). Psychology of work behavior (4th edition). Pacific Grove, CA: Brooks/Cole. Chapters 4 and 5.

Landy, F. J., & Farr, J. L. (1980). Performance ratings. Psychological Bulletin, 87, 72-107.

Lane, J., & Herriot, P. (1990). Self-ratings, supervisor ratings, positions and performance. Journal of Occupational Psychology, 63, 77-88.

Latham, G. P., & Locke E. A. (1991). Self-regulation through goal-setting. Organizational Behavior and Human Decision Processes, 50, 212-247.

Latham, G. P., & Wexley, K. N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.

Levine, J. M., & Moreland, R. L. (1986). Outcome comparisons in group contexts: Consequences for the self and others. In R. Schwarzer (Ed.), Self-related cognitions in anxiety and motivation, (pp. 285-303), Hillsdale, N.J.: Erlbaum.

Levine, J. M., & Moreland, R. L. (1987). Social comparison and outcome evaluation in group contexts. In J. C. Masters & W. P. Smith (Eds.), Social comparison, social justice, and relative deprivation: Theoretical, empirical, and policy perspectives (pp. 105-127). Hillsdale, NJ: Erlbaum.

London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. Personnel Psychology, 44, 375-390.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. Academy of Management Review, 1, 183-193.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. Journal of Applied Psychology, 67(3), 280-296.

McDonald, T. (1991). The effect of dimension content on observation and ratings of job performance. Organizational Behavior and Human Decision Processes, 48, 252-271.

McEnery, J., & McEnery, J. M. (1987). Self-ratings in management training needs assessment: A neglected opportunity? Journal of Occupational Psychology, 60, 49-60.

McEvoy, G. M., & Buller, P. F. (1987). User acceptance of peer appraisals in an industrial setting. Personnel Psychology, 40, 785-797.

Meyer, H. (1980). Self-appraisal of job performance. Personnel Psychology, 33, 291-295.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. Personnel Psychology, 37, 687-702.

Mowday, R. T. (1987). Equity theory predictions of behavior in organizations. In R. M. Steers & L. W. Porter (Eds.), Motivation and Work Behavior (4th ed.), (pp. 89-110). New York: McGraw-Hill, Inc.

Murphy, K. R., & Cleveland, J. N. (1991). Performance appraisal: An organizational perspective. Boston, MA: Allyn and Bacon.

Myers, D. G. (1992). Psychology (3rd ed.). New York: Worth Publishers, Inc.

Nathan, B. R., & Tippins, N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. Journal of Applied Psychology, 75, 290-296.

Oldham, G. R., Kulik, C. T., Ambrose, M. L., Stepina, L. P., & Brand, J. F. (1986). Relations between job facet comparisons and employee relations. Organizational Behavior and Human Decisions Processes, 38, 28-47.

Parker, J. W., Taylor, E. K., Barrett, R. S., & Martins, L. (1959). Rating scale content III: Relationship between supervisory and self-ratings. Personnel Psychology, 12, 49-63.

Pulakos, E. D. (1984). The development of training rater programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decision Processes, 38, 76-91.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Riggio, R. E., & Cole, E. J. (1992). Agreement between subordinate and superior ratings of supervisory performance and effects on self and subordinate job satisfaction. Journal of Occupational and Organizational Psychology, 65, 151-158.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessments. Psychological Bulletin, 90, 322-351.

Smith, D. E. (1986). Training programs for performance appraisal: A review. Academy of Management Review, 11, 22-40.

Somers, M. J., & Birnbaum, D. (1991). Assessing self-appraisal of job performance as an evaluation device: Are the poor results a function of method or methodology? Human Relations, 44, 1081-1091.

Steel, R. P., & Ovalle, N. K. (1984). Self-appraisal based upon supervisory feedback. Personnel Psychology, 37, 667-684.

Stepina, L. P., & Perrewe, P. L. (1991). The stability of comparative referent choice and feelings of inequity: A longitudinal field study. Journal of Organizational Behavior, 12, 185-200.

Suls, J., & Wills, T. A. (1991). Social comparison: Contemporary theory and research (Eds.), Hillsdale, N.J.: Erlbaum.

Sulsky, L. & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. Journal of Applied Psychology, 77, 501-510.

Summers, T. P., & DeNisi, A. S. (1990). In search of adams' other: Reexamination of referents used in the evaluation of pay. Human Relations, 43, 497-511.

Sweeney, P. D., McFarlin, D. B., & Inderrieden, E. J. (1990). Using relative deprivation theory to explain satisfaction with income and pay level: A multistudy examination. Academy of Management Journal, 33, 423-436.

Thornton, G. C. (1968). The relationship between supervisory and self- appraisals of executive performance. Personnel Psychology, 21, 441-455.

Thornton, G. C. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-271.

Weiner, B. (1986). An attributional theory of motivation and emotion. New York: Springer-Verlag.

APPENDIX A

PACKET INSTRUCTIONS AND INFORMED CONSENT SHEET

FOR SELF-RATERS

_____
Your Name

_____
Supervisor's Name

PERFORMANCE APPRAISAL QUESTIONS AND RATINGS PACKET

You are going to be asked to fill out a series of questions and rating scales pertaining to your job performance and on what basis you evaluate your performance. It is VERY IMPORTANT that you DO NOT look ahead; proceed one page at a time. Please provide honest and accurate responses for all questions and ratings. The entire packet should take approximately 5 minutes to complete.

All responses made in this packet will remain confidential and will be used for research purposes ONLY. Your individual responses WILL NOT be made available to your supervisor, your coworkers, or your company. The purpose of listing you and your supervisor's name at the top of this sheet is only to ensure that employees and supervisors can be matched together for research purposes.

By signing and dating this form you are providing your voluntary consent to participate in this research (by filling out the remainder of the packet) as described above.

_____ _____
Signature Date

FOR SUPERVISORS

Your Name

Employee's Name

PERFORMANCE APPRAISAL QUESTIONS AND RATINGS PACKET

You are going to be asked to fill out a series of questions and rating scales pertaining to your employee's job performance and on what basis you evaluate their performance. It is VERY IMPORTANT that you DO NOT look ahead; proceed one page at a time. Please provide honest and accurate responses for all questions and ratings. The entire packet should take approximately 5 minutes to complete.

All responses made in this packet will remain confidential and will be used for research purposes ONLY. Your individual responses WILL NOT be made available to your employees or your company. The purpose of listing you and the employee's name at the top of this sheet is only to ensure that employees and supervisors can be matched together for research purposes.

By signing and dating this form you are providing your voluntary consent to participate in this research (by filling out the remainder of the packet) as described above.

Signature Date

APPENDIX B

PRE-RATING COMPARISON STANDARD QUESTIONS

FOR SELF-RATERS

Please think about how you would rate your own job performance. If asked to evaluate your own performance on the job (i.e., provide a self-rating) what would you use as the basis for your ratings? That is, how would you decide whether or not you were performing satisfactorily on the job?

Please answer in your own words

FOR SUPERVISORS

Please think about how you would (or do) rate your employee's job performance. If asked to rate an employee on his/her job performance (i.e., provide a supervisory rating), what would you use as the basis for your ratings? That is, how would you decide whether or not the employee was performing satisfactorily on the job?

Please answer in your own words

APPENDIX C

RATING INSTRUCTIONS AND PERFORMANCE DIMENSIONS

RATING SHEET INSTRUCTIONS

The next five pages will be asking you to make ratings across three different performance dimensions. The five rating sheets are exactly identical EXCEPT for the instructions on how to generate your ratings. It is VERY IMPORTANT that you read the instructions at the top of each page carefully and provide ratings in a manner consistent with the specific instructions. Listed below are the definitions of what constitutes a specific performance dimension for each of the three dimensions.

DIMENSION 1

DIMENSION 2

FOR SUPERVISOR

OVERALL PERFORMANCE - The overall job performance level of the employee when considering both of the previous dimensions together.

FOR SELF-RATER

OVERALL PERFORMANCE - Your overall job performance level when considering both of the previous dimensions together.

APPENDIX D

SELF-EVALUATION RATING SHEETS

AMBIGUOUS[4]

Based on your performance over the past six months, please rate yourself on the following performance dimensions.

Please circle the appropriate number for each dimension

DIMENSION 1

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor       Poor         Average       Good       Very Good
```

DIMENSION 2

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor       Poor         Average       Good       Very Good
```

OVERALL PERFORMANCE

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor       Poor         Average       Good       Very Good
```

---

[4]The title "AMBIGUOUS" will not actually be used on the subject's rating sheet but will instead be left blank.

INTERNAL

Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your own personal, internal values and standards as a criteria. That is, base your ratings on how well you personally feel you have done over the past six months relative to your abilities and past performance. DO NOT give consideration to any other criteria beyond your own beliefs as to how well you performed.

Please circle the appropriate number for each dimension

DIMENSION 1

1 2 3 4 5 6 7 8 9

Very Poor Poor Average Good Very Good

DIMENSION 2

1 2 3 4 5 6 7 8 9

Very Poor Poor Average Good Very Good

OVERALL PERFORMANCE

1 2 3 4 5 6 7 8 9

Very Poor Poor Average Good Very Good

ABSOLUTE

Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your company's minimum requirement or goal as a criteria. That is, for each dimension rate yourself in comparison to the minimal level of performance as defined by your company or group's policy. DO NOT give consideration to any other criteria beyond your own belief as to **whether or not you met this requirement.**

Please circle the appropriate number for each dimension

DIMENSION 1

1	2	3	4	5	6	7	8	9
Very Poor		Poor		Average		Good		Very Good

DIMENSION 2

1	2	3	4	5	6	7	8	9
Very Poor		Poor		Average		Good		Very Good

OVERALL PERFORMANCE

1	2	3	4	5	6	7	8	9
Very Poor		Poor		Average		Good		Very Good

RELATIVE

Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your fellow coworkers' performance as a criteria. That is, think about how your co-workers have performed and compare yourself to them. DO NOT give consideration to any other criteria beyond your own belief as to how well you performed **in direct comparison to your co-workers.**

Please circle the appropriate number for each dimension

DIMENSION 1

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor      Poor         Average       Good       Very Good
```

DIMENSION 2

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor      Poor         Average       Good       Very Good
```

OVERALL PERFORMANCE

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor      Poor         Average       Good       Very Good
```

MULTIPLE

Based on your performance over the past six months, please rate yourself on the following performance dimensions. Use your own personal standards, your attainment of the minimum requirements and goals, and your comparison with fellow co-workers as the criteria. That is, consider all three standards as defined in the previous pages. Give equal consideration to all three of the criteria.

Please circle the appropriate number for each dimension

DIMENSION 1

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor      Poor         Average       Good       Very Good
```

DIMENSION 2

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor      Poor         Average       Good       Very Good
```

OVERALL PERFORMANCE

```
    1      2      3      4      5      6      7      8      9
    *------*------*------*------*------*------*------*------*
Very Poor      Poor         Average       Good       Very Good
```

APPENDIX E

SUPERVISORY RATING SHEETS

AMBIGUOUS

Based on your employee's performance over the past six months, please rate this employee on the following performance dimensions.

INTERNAL

Based on your employee's performance over the past six months, please rate this employee on the following performance dimensions. Use your perceptions of the employee's own personal, internal values and standards as a criteria. That is, base your ratings on how you think the employee feels they have done over the past six months relative to their abilities and past performance. DO NOT give consideration to any other criteria beyond how **you believe the employee perceives they have done** over the past six months.

ABSOLUTE

Based on your employee's performance over the past six months, please rate this employee on the following performance dimensions. Use your company's minimum requirement or goal as the criterion. That is, for each dimension rate the employee in comparison to the minimal level of performance as defined by your company or department's policy. DO NOT give consideration to any other criteria beyond your own belief as to **whether or not the employee met this requirement.**

RELATIVE

Based on your employee's performance over the past six months, please rate this employee on the following performance dimensions. Use the employee's fellow coworkers' performance as a criteria. That is, think about how the employee's co-workers have performed and compare the employee to them. DO NOT give consideration to any other criteria beyond your own belief as to how well the employee performed **in direct comparison to his/her other co-workers.**

(appendix continued)

APPENDIX E (con'd)

MULTIPLE

Based on your employee's performance over the past six months, please rate this employee on the following performance dimensions. Use your perceptions of the employee's own personal standards, the employee's attainment of the minimum requirements and goals, AND comparison of the employee's performance with fellow co-workers as the criteria. That is, consider all three standards as defined in the previous pages. Give equal consideration to all three of the criteria.

APPENDIX F

POST-RATING COMPARISON STANDARD QUESTIONS

FOR SELF-RATERS

If asked to evaluate your own performance in the future, please rate each of the four comparison standards as to your preference for using them in future performance ratings.

You may refer back to the comparison standard instructions on the previous rating sheets if you need to.

Please circle the appropriate number for each standard

INTERNAL Standard (Own internal values and standards)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

ABSOLUTE Standard (Company's min. requirement/goal)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

RELATIVE Standard (Performance of fellow co-workers)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

MULTIPLE Standard (Combination of previous standards)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

FOR SUPERVISORS

If asked to rate employees in the future, please rate each of the four comparison standards as to your preference for using them in future performance appraisals.

You may refer back to the comparison standard instructions on the previous rating sheets if you need to.

Please circle the appropriate number for each standard

INTERNAL Standard (Own internal values and standards)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

ABSOLUTE Standard (Company's min. requirement/goal)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

RELATIVE Standard (Performance of fellow co-workers)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

MULTIPLE Standard (Combination of previous standards)

1	2	3	4	5	6	7	8	9
Very Low Preference		Low Preference		Neutral		High Preference		Very High Preference

APPENDIX G

AVAILABILITY RATINGS

AVAILABILITY OF COMPARISON STANDARD INFORMATION

Please rate the degree to which information for each comparison standard was available to help you make performance ratings. That is, to what extent was information for each comparison standard readily and easily obtained within your own workplace.

You may refer back to the comparison standard instructions on the previous rating sheets if you need to.

Please circle the appropriate number for each standard

INTERNAL Standard (Own internal values and standards)

1	2	3	4	5
Not Available		Moderately Available		Very Available

ABSOLUTE Standard (Company's min. requirement/goal)

1	2	3	4	5
Not Available		Moderately Available		Very Available

RELATIVE Standard (Performance of fellow co-workers)

1	2	3	4	5
Not Available		Moderately Available		Very Available

MULTIPLE Standard (Combination of previous standards)

1	2	3	4	5
Not Available		Moderately Available		Very Available

APPENDIX H

RELEVANCY RATINGS

RELEVANCY OF COMPARISON STANDARD INFORMATION

Please rate the degree to which each comparison standard was pertinent and relevant to base your performance ratings on. That is, to what extent were the comparison standards important and applicable within your own workplace as a basis for your ratings.

You may refer back to the comparison standard instructions on the previous rating sheets if you need to.

Please circle the appropriate number for each standard

INTERNAL Standard (Own internal values and standards)

1	2	3	4	5
Not Relevant		Moderately Relevant		Very Relevant

ABSOLUTE Standard (Company's min. requirement/goal)

1	2	3	4	5
Not Relevant		Moderately Relevant		Very Relevant

RELATIVE Standard (Performance of fellow co-workers)

1	2	3	4	5
Not Relevant		Moderately Relevant		Very Relevant

MULTIPLE Standard (Combination of previous standards)

1	2	3	4	5
Not Relevant		Moderately Relevant		Very Relevant

APPENDIX I

DEMOGRAPHICS AND COMPREHENSION QUESTION

PERSONAL INFORMATION

Age: _____ Sex: _____

Current Job Title/Occupation: _____

Number of Years with Company: _____

Number of Years at Present Job/Position: _____

(FOR SUPERVISORS)
Number of Employees under your direct supervision: _____

Do you feel you understood all the instructions and questions asked throughout this packet and were able to answer them in an honest and accurate manner? Y or N

VITA

Brian Wayne Schrader was born on December 19th, 1965 in Sioux City, Iowa. He graduated from Andover High School in Andover, Kansas in May of 1984. In May of 1988, he graduated from Bethany College, located in Lindsborg, Kansas, with a Bachelor of Arts degree in both Psychology and Chemistry. In December of 1990, he graduated with a Master of Arts degree in Psychology from Louisiana State University located in Baton Rouge, Louisiana.

He completed his Doctor of Philosophy degree in Industrial/Organizational Psychology from Louisiana State University and will receive it at the fall commencement of 1993.

Dr. Schrader is currently an Assistant Professor in Psychology at St. Xavier University in Chicago, Illinois. He began this tenure-track position in the fall of 1993.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Brian W. Schrader

Major Field: Psychology

Title of Dissertation: Differential Comparison Standards and Their Subsequent Effects on the Agreement Between Self- and Supervisor Performance Appraisal Ratings.

Approved:

Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

Date of Examination:

October 22, 1993